The Dissertation Committee for Anthony James Bendinelli

certifies that this is the approved version of the following dissertation:

The Application of Visualization Methods to Educational Data Sets with Inspiration from Statistical and Fluid Mechanics

Committee:

Michael Marder, Supervisor

Philip Morrison

Qian Niu

Harry Swinney

Jill Marshall

The Application of Visualization Methods to Educational Data Sets with Inspiration from Statistical and Fluid Mechanics

by

Anthony James Bendinelli, B.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2014

Dedicated to my family, for never failing to believe in me

Acknowledgments

The proverb states that it takes a village to raise a child. It is also true that it takes a community to raise a scientist. I would not be here today without the support and help of many, many people.

I would like to thank Dr. Michael Marder for taking me under his wing and helping me develop as a researcher and as a scientist. He held our work to the highest standards and proved to be a great inspiration in the process. Mike's boundless enthusiasm for our research (and for everything else!) made for wonderfully enjoyable conversations. He is a great mentor and a great friend, and an incredibly valuable member of the physics community.

I would like to thank Greg Branch, Leigh Hausman, and the staffs of the Educational Research Centers in both Austin and Dallas. We could not have conducted this research without their help and understanding. Our research methods were quite unusual, but they bent over backwards to help us in any way they could.

I would like to thank Olga Vera, Marybeth Casias, and Rachael Salge for keeping the Center of Nonlinear Dynamics afloat and helping us work through every administrative hurdle. Without their support and hard work, the Center would be in much worse shape.

I would like to thank Dhruv Bansal, David McGhan, and Matthew Guthrie. Dhruv laid the groundwork for this research before moving on to other opportunities in the private sector; he kept encouraging me to become a better programmer even when it seemed that nothing was working. Dave and Matt are great coworkers and great friends that will carry the torch forward, and hopefully discover even more exciting results from this research.

I would like to thank each current and past member of the Center for Nonlinear Dynamics, who are not only fantastic researchers but also awesome people. The camaraderie that we developed helped us all to push through the difficulties inherent in graduate school. I would like to especially thank Frank Male, Chih-Hung Chen, and Andrea Keidel for being pillars of support during the last two years (as well as for countless hours playing basketball!).

I would like to thank my entire family for supporting me throughout the process of graduate school. My parents, Chris and Linda, and my sister Stephanie were always willing to talk with me, laugh with me, cheer me up, listen to my frustrations, and encourage me every step of the way. My extended family also kept me sane and supported me; in particular, I would like to thank my grandparents James, Lois, Richard, and Norma Jean. Without my family, I would not be where I am today.

Finally, I would like to extend my heartfelt thanks to Tiffany Gatchel, who has been by my side ever since I arrived in Austin seven years ago. Through all the ups and downs, she has never wavered in her love and support. Tiffany and her family have become *my* family in Texas, and I could not have

done this without them. Thank you, Tiffany, for all that you are and all that you have done.

The Application of Visualization Methods to Educational Data Sets with Inspiration from Statistical and Fluid Mechanics

Anthony James Bendinelli, Ph.D.
The University of Texas at Austin, 2014

Supervisor: Michael Marder

This dissertation focuses on the development of visualization methods that enable us to examine longitudinal data in a unique way. We take inspiration from statistical and fluid mechanics to represent our data as a "flow" through time. Our visualizations represent vector fields (or *flow plots*), streamlines, and trajectories, and they are constructed in a similar manner to how one might analyze the aggregate motion of particles in a fluid.

However, the subject of our research extends beyond ordinary fluid mechanics. We will use our visualizations to examine statewide standardized test scores in Texas from 2003 to 2011. The nature of the data makes it a perfect match for our methodology, since students' test scores tend to change over time in a semi-deterministic but nonlinear manner. Furthermore, our methods represent a departure from the standard ways of analyzing educational data.

By visualizing the changes in students' test scores over a nine-year period, we discovered that our flow plots were changing with the eventual graduating class of 2012. The change in our visualizations was caused by an educational policy known as the Student Success Initiative, or SSI. The policy forced students to pass their standardized tests in 5th and 8th grade, or risk being held back a grade. To help with this process, students who initially failed were given extra instruction and additional opportunities to take the test. SSI was implemented in such a way that it would affect the class of 2012 and beyond, although we did not know of the program's existence until our plots had been developed.

SSI had a successful impact on the educational career of Texas students; a far greater percentage of students were able to pass the 5th and 8th grade standardized tests after SSI was implemented. The striking feature of SSI, however, is that it also significantly improved test scores in 6th, 7th, 9th, and 10th grade. Despite its success at improving test scores over many years and grades, the program was eventually defunded. This was partially due to an inability to construct a lengthy longitudinal analysis of the program's influence.

Our methodology would have conclusively shown the effectiveness of the SSI policy. Despite the defunding of the SSI, I am confident our methodology can be extended to illustrate changes in other data systems. These systems may or may not be related to education; our code and techniques are designed to be as universal as possible. We will explore several extensions to other data

sets at the end of this dissertation.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Students take standardized tests throughout their school careers, and in the process they generate large amounts of data that can prove quite valuable to educational researchers. Many statistical methods have been developed to analyze data of this nature. This thesis describes a method for analysis that strongly emphasizes communication through visualization.

The initial basis for this research is a method for analyzing educational data introduced in Marder and Bansal [31]. In that paper, the authors took techniques for describing convective and diffusive particle flows and applied them to mathematics scores on standardized tests. I have extended and refined the concepts in that paper, borrowing additional terminology and techniques from statistical and fluid mechanics. I have created new visualization methods that track the progress of specific groups of students throughout their educational careers. These methods are unorthodox in an educational context, but they are powerful in their ability to communicate results to people regardless of technical background.

The visualizations are created using data collected from Texas' stan-

dardized tests. During the course of my research, I discovered that the visualizations were changing in dramatic ways beginning in 2005. These changes turned out to be the fingerprint of a statewide educational policy which has since been defunded. My belief is that these techniques can be applied to other state- and national-level data, and identify the successes (or failures) of major policy changes.

This thesis will focus entirely on educational data, but the methods described herein may be extended to other longitudinal data sets. It is important to note that this research is different from what is traditionally called "physics education research", which is the study of how physics may be taught better at every level of school. Instead, this research focuses on using physics (and the training/intuition that comes from studying it) to analyze data and answer questions in the field of educational research. Physicists may have an easier time understanding the methods/techniques and their derivations, but the results are of interest to people involved in education.

1.1 "No Child Left Behind" Act

The Elementary and Secondary Education Act of 2001 (No Child Left Behind, or NCLB) was intended to increase the accountability of state education systems with regard to the progress of their students [2]. States were required to adopt academic standards and create a series of standardized tests that would evaluate the students' progress towards meeting those standards.

	AYP Performance Standards for 2002-03 to 2013-14				
School Year(s)	2002-04	2004-06	2006-08	2008-09	2009-10
Reading/English Language Arts	47%	53%	60%	67%	73%
Mathematics	33%	42%	50%	58%	67%
School Year(s)	2010-11	2011-12	2012-13	2013-14	—
Reading/English Language Arts	80%	87%	93%	100%	—
Mathematics	75%	83%	92%	100%	—

Table 1.1: List of AYP Performance Standards in Texas by School Year [48]

Each state was allowed to create its own version(s) of standardized tests, so the resulting test scores of Ohio's students might not be comparable to the results of Oregon's students. Furthermore, each state was allowed to select the materials that would be covered/emphasized on these standardized tests. Schools are evaluated based on their "adequate yearly progress" (AYP) towards meeting benchmarks established by the federal government (Table 1.1). Each school is required to report a passing rate above the AYP benchmark for that particular year, and there are separate annual objectives for certain subgroups of students, including economically disadvantaged students, students from major ethnic groups, students with disabilities, and students with limited English proficiency [2].

Schools that fail to meet these AYP benchmarks for two consecutive years are required to undergo some form of rehabilitation. NCLB offers some suggestions (including allowing students to transfer to other public schools or "decreas[ing] management authority at the school level" [2]) but it is ulti-

mately up to the local educational agency to determine how rehabilitation is to proceed.

Texas has a five-stage plan for rehabilitation [47]. In the first year, schools must submit a Campus Improvement Plan to devise strategies to improve performance and meet AYP standards. In the second year of non-compliance, schools are required to provide supplemental educational services, such as tutoring, although additional funds are not given to the local education agencies to cope with these additional services. In the third year, schools are marked for "corrective action" and the local educational agency must implement at least one of the following options:

- Replace the school staff
- Institute a new curriculum
- Decrease management authority at the school level
- Appoint an outside expert to advise the campus on its progress
- Extend the school year or school day
- Restructure the internal organizational structure of the school

In the fourth year, the local educational agency increases its oversight over the underperforming campus, and it plans for a complete restructuring of the school if it cannot turn its performance around. Finally, in the fifth year the school staff are replaced and the operation of the school may be turned over

to the State of Texas. The school may also be reopened as a charter school [47].

As of December 2012, 1154 of the 8529 schools across the state of Texas have not met AYP benchmarks for at least two years and are marked for some form of rehabilitation [9].[1] Of those, 971 schools are in Stage 1 of rehabilitation, while 59 schools are in Stage 5 and marked for imminent closure [8]. Furthermore, the number of schools that failed to meet AYP standards is on the rise. In 2009 and 2010, only 353 and 368 schools in the state of Texas underperformed on standardized tests. However, in 2011 the number of underperforming schools increased to 2190, and in 2012 there were 4054 schools (48% of the total schools in Texas) that failed to reach the AYP benchmarks [8].

This spike in the number of underperforming schools is alarming, but not necessarily surprising. As shown in Table 1.1, the standards required by the federal government increase every year. It seems, though, that the federal standards were assembled without considering how education works or even simple statistical truths. The end goal of NCLB was to raise standards to the point where all students have a basic understanding of core concepts as determined by each state. But to require a 100% passing rate on a test is to expect the impossible. Expecting a passing rate above 95% is to expect the near-impossible. Some schools will be able to achieve these high standards,

[1] More recent information about AYP compliance is not available from official websites.

and they are to be commended for their performance. To demand these standards for all students nationwide is quite unreasonable. To illustrate the point further, if the standardized test was one question that simply said "Spell your name correctly", there still would not be 100% passing rates across the entire state.

As schools struggle to meet the AYP benchmarks, the explicit assumption is that, if students are underperforming on these standardized tests, the teachers should be replaced with better teachers and then the students will be able to meet standards.[2] Whether or not this would actually occur seems unclear. Hanushek and Rivkin estimate that, if a student had a teacher from the 75th percentile of teaching quality instead of a teacher from the 25th percentile, the student would experience a gain of 0.2 standard deviations in a single year [26]. In Texas, where the standard deviations of scores on the standardized math test are on the order of 15-20%, this gain would be equivalent to scoring 3-4% higher. The quality of teachers is not insignificant when determining how much students learn and how well they perform on tests. However, if a student scores 45% on a standardized test and needs a 60% to pass, will firing the teachers and replacing them with better ones make up that 15% gap? It might work in some cases and not work in others, but there can be no margin for error when the state demands passing rates above 90%.

[2]As of March 2014, 45 states have applied for waivers to circumvent this process by requesting flexibility from the requirements of NCLB [1].

1.1.1 Value-Added Assessments

There are some who believe that it is possible to evaluate the effects that teachers have on students in a way that is fair, unbiased, and independent of any confounding factors such as parental influence or availability of learning materials.. In 1993, William Sanders and Sandra Horn developed the Tennessee Value-Added Assessment System (TVAAS) to evaluate teachers based on how much their students learn over a given year, as opposed to whether they meet the requirements of a single test [43]. Any outside influences that would affect student performance are accounted for by having each student serve as his or her own "control" [43]. Sanders declares his personal belief that "any of these [value-added] models should not include socio-economic or ethnic accommodations but should only include measures of previous achievement of individual students" [42].

The TVAAS showed great promise in assessing teacher effects. In 1996, Sanders and Rivers showed that teachers had extreme impacts on students between second and fifth grade. They analyzed two major metropolitan school systems in Tennessee by dividing teachers into quintiles based on their teaching ability. They showed that fifth-grade students who had received three years of instruction from teachers in the top quintile tended to have a mean student percentile that was 50 points above those fifth-graders who had been taught for three years by teachers in the bottom quintile [41]. In 1997, Wright, Horn, and Sanders used a mixed-model approach to estimate how a student's gain score is affected by the teacher, the school system, the class size, the achievement

level of the student, and the heterogeneity of the classroom with respect to achievement levels. By analyzing the z-scores for gains between third and fifth grade, they concluded that teachers and achievement level are the two most important factors to impact student gain [52]. These studies form part of the basis for Sanders' claim that "differences in teacher effectiveness is [sic] the single largest factor affecting academic growth of populations of students" [42].

Gordon, Kane, and Staiger (2006) published a report that used value-added assessments to see whether teacher certification is indicative of teacher effectiveness [25]. After controlling for demographic characteristics of students, the authors found that teacher certification had little to no impact on student achievement, but the average difference between a teacher in the top quartile of effectiveness and one in the bottom quartile is worth 10 percentile points. They extrapolate this finding to claim that, if the effects were cumulative, a student that receives four years of top-quartile instruction (instead of bottom-quartile) could cover the national black-white achievement gap of 34 percentile points. The authors then propose several measures to improve the overall effectiveness of teachers which include reducing the barriers that limit non-certified teachers from teaching, making it harder to give tenure to the least effective teachers, and incentivizing high-quality teachers with financial bonuses if they agree to teach at poorer schools. They estimate that denying tenure to the least effective teachers could potentially be worth \$216 billion to \$507 billion per year in terms of the nationwide economic value of raising stu-

dents' academic test scores [25]. This report has been cited in the Los Angeles Times' decision to release the value-added ratings of all Los Angeles Unified School District elementary school teachers [19], and its influence can be seen in President Barack Obama's Race to the Top program that asks states to improve "teacher and principal effectiveness based on performance" and ensure "equitable distribution of effective teachers and principals" [35].

However, several studies have contested the results listed above. McCaffrey et al. reviewed the literature and specifically analyzed Sanders and Rivers (1996) and Wright, Horn, and Sanders (1997) due to those being two of the most widely cited papers to use value-added methods [33]. For Sanders and Rivers, they note that the same students are used to estimate teacher quintiles *and* deduce the quintiles' effects on student performance; their conclusion is that there is evidence that teachers affect student performance, but the size of the effect is not nearly as large as what was reported [33]. In the case of Wright, Horn, and Sanders, they disregard the reported z-scores as useful indicators of effect sizes because they are depend on details of the analysis such as sample size, correlations among predictors, and the effect sizes themselves [33]. Finally, Marder and Bansal's model showed that score gains depend heavily on socio-economic status when examined on a statewide level [31]. On average, individuals who receive free or reduced-price lunches will have lower score gains than those who do not, despite having the same initial test scores. This result may be reconciled with Sanders' claims if the quality of teachers is significantly lower for low-income students than for better-off

students; however, it is difficult to test this hypothesis without risking the circular logic that ineffective teachers are defined to be those for whom students' score gains are lower.

The report of Gordon, Kane, and Staiger contains challengeable assertions as well. First, the conclusions are drawn from the test scores of third through fifth graders from the Los Angeles Unified School District. It may not be valid to use conclusions about the effectiveness of teachers in elementary school and extrapolate them to middle or high school (see Section 2.4). Second, the study controls for baseline characteristics of students and prior year scores by using linear regression. However, certain socio-economic variables may have a nonlinear effect on student test scores, such as whether a student receives free/reduced-price lunches or not ([31, 15]; see Section 4.3.2). Finally, it is unclear how the authors arrived at some of their conclusions. They estimate that replacing ineffective teachers with novice teachers would produce student test score gains of 1.2 percentile points per year, but the formula used to calculate this gain includes values that are not found anywhere else in their report. Furthermore, they make the assumption that the pool of novice teachers would have a similar distribution of teaching effectiveness to the current set of Los Angeles elementary school teachers; they do not consider that their policy suggestions might alter the overall quality of the incoming teachers.

The problem of erroneous extrapolation is not unique to Gordon, Kane, and Staiger. Hoxby, Murarka, and Kang (2009) conducted an evaluation of New York City's charter schools to see how they affected their students'

achievement on tests [28]. The research does not depend on value-added measurements to compare charter school teachers to other public school teachers. Instead, it compares the population of students who were randomly selected from a lottery to attend charter schools to the population of students who entered the lottery but did not "win". This study has the benefit of comparing two populations which are similar in demographics and attitudes towards education, but were randomly assigned to different schools. Hoxby et al. found that the students who attended charter schools from kindergarten through 8th grade could overcome 86 percent of the "Scarsdale-Harlem achievement gap" in mathematics, defined to be 35 scale score points or approximately the difference between the levels of "not meeting learning standards" and "meeting learning standards."[3] This calculation comes from charter school students being 0.14 standard scores ahead of non-charter school students by 3rd grade, and then outpacing non-charter school students by 0.12 standard scores more for each year after that.

However, in a technical paper released later, the authors admit that only 25% of their student population had received six or more years of charter school education [27]. Hoxby et al. estimated a single year gain and multiplied it over the course of several years to reach their conclusions. Furthermore, only 40% of the schools at the time of the study had been open for six or more years. While the average annual gain is not necessarily an incorrect estimate,

[3]The name "Scarsdale-Harlem" refers to the approximate 35-point achievement gap between Harlem and Scarsdale, an affluent New York City suburb [28]. Many of New York City's charter schools are located in Harlem.

the linear extrapolation over a nine-year span (for which the researchers did not have a full nine years of data) creates misleading conclusions [20].

While AYP may seem like a type of value-added system, schools are only evaluated based on either the absolute number of students who pass the test, or by a relative decrease in the number of students who fail. In Texas, if there is a 10% decrease in the number of failing students, that is sufficient to show Adequate Yearly Progress (along with attaining certain benchmarks relating to graduation/attendance rates) [48]; the contributions of a teacher who raises a student from the 20th percentile to the 50th percentile are effectively ignored, and so AYP cannot be considered as any type of value-added assessment. Furthermore, the AYP requirements that are applied to all students are also applied, individually, to the African American, Hispanic, White, economically disadvantaged, special education, and limited English proficient student groups [48]. No accommodations are made for different ethnic or socio-economic groups, despite the fact that score gains are strongly influenced by socio-economic factors [42, 16, 31].

1.2 Texas Assessment of Knowledge and Skills (TAKS)

As mentioned above, all states are required to create some kind of standardized test in order to evaluate AYP, and each state controls and develops its own standards. It is possible for the Iowa standardized test to cover drastically different material than the Texas standardized test [50]. Therefore,

attempting to draw comparisons between states is difficult, if not impossible, without the aid of national tests/assessments like the National Assessment of Educational Progress (NAEP) or the SAT tests.

The Texas Assessment of Knowledge and Skills (TAKS) is the standardized test used by Texas since the 2002-03 school year. It was created after a three-year development process to determine a standardized test that would assess students' understanding of the material required by the Texas state standards, also known as the Texas Essential Knowledge and Skills (TEKS) [12, 13]. TAKS was not the first standardized test used by Texas; the Texas Assessment of Academic Skills (TAAS) had been used by the state since 1990 [5]. TAKS was intended to be an updated version of TAAS that would serve as a more authentic indicator for students' understanding of TEKS material. In the 2011-2012 school year, TAKS was again updated to the State of Texas Assessments of Academic Readiness (STAAR). STAAR differs from TAKS mainly at the high school level, where generic tests of mathematics or science are replaced with subject-specific tests like geometry, algebra II, and chemistry [11]. This research focuses exclusively on TAKS; possible extensions to STAAR or TAAS will be elaborated on later (see Section 5.1).

Standardized tests (and tests in general) have been around for decades. The effect of NCLB in 2002, however, was that states were now recording the test results of millions of students across every state, and many were allowing researchers to utilize and explore that data. The Texas Schools Project (TSP) through the University of Texas at Dallas is one such research organization.

	Reading	Writing	Math	Science	Soc. Studies
Grade 3	X		X		
Grade 4	X	X	X		
Grade 5	X		X	X	
Grade 6	X		X		
Grade 7	X	X	X		
Grade 8	X		X	X	X
Grade 9	X		X		
Grade 10	X		X	X	X
Grade 11	X		X	X	X

Table 1.2: TAKS subject tests by grade level

The TSP has a contract with the Texas Education Agency (TEA) that grants them access to this wealth of standardized test data; in turn, the TSP and TEA may also grant access to outside researchers. Without this access, the Marder and Bansal methodology would have been impossible to test and verify.

The TAKS test was given to all public school students between third grade and eleventh grade from 2003 to 2011. It covers the subjects of reading, writing, mathematics, science, and social studies, although not all subjects are tested at each grade (see Table 1.2). In 10th and 11th grade, reading and writing are combined together and are tested as English Language Arts. Mathematics is the only subject to be tested individually and at every grade level, making it the easiest subject for one to assemble a longitudinal data record. Also, the number of questions asked on each grade's test has remained the same since the inception of TAKS, although the number of questions needed to pass has not.

This leads to questions about how students' scores may be reported. The most trivial calculation is to take scores simply as a *raw percentage* corresponding to the number of questions correctly answered. The TEA also provides *scaled scores* that attempt to use a common scale which accounts for the difficulty of a single assessment. In the TEA's scaled score format, a 2100 corresponds to a passing score, and a 2400 indicates Commended Performance. The number of questions needed to reach the levels of 2100 or 2400 was determined by a panel of experts, who also judged the test questions for content validity and grade appropriateness. By assigning difficulty levels to each question, it became possible to create exams of roughly equal difficulty from one year to the next. Also, there was very little deviation in the number of raw questions needed to pass the test from year to year; usually the passing benchmark varies by only one or two questions. The precise process that the panel used to create these scaled scores is proprietary, so the exact procedure is unknown. The effect of variations in passing scores will be discussed later in the paper (see Section 1.6). For now, it is sufficient to note that the 2100/2400 benchmarks were in place every year until the TEA stopped reporting scaled scores in 2010.[4]

In 2007, the TEA began reporting *Quantile measures* in addition to its

[4]The passing scores in 2003 and 2004 have lower scaled scores for passing, but that was due to a two-year ramp in which the required benchmark was gradually raised until it met with the panel's recommendation for 2100.

scaled scores. Quantile measures on math exams (and their Lexile measure counterparts for reading exams) are a measure of test difficulty created by MetaMetrics, a private company. Many states use this additional metric, but it is only included here for completeness, as it seems that Quantile measures do not include any sort of qualitative analysis of test difficulty.

In 2009, the TEA created *vertical scaled scores*. These scores work on a 0 to 1000 scale, and vertical scaled scores should allow comparisons between different test grades for the same subject. The typical progress that a student makes from one year to the next is defined as the difference between passing standard cut scores. As an example, if the 5th grade mathematics test has a passing cutoff of 603 and the 6th grade mathematics test has a passing cutoff of 637, then an average student is expected to gain 34 points from 5th grade to 6th grade. A student's vertical scaled scores may be evaluated against these standards to see if the student is ahead or behind the progress rate of others, and allow teachers to compensate accordingly. While this measure could be useful for evaluating students longitudinally, it was not created until late in the TAKS's lifecycle, and the process by which raw percentages are converted to vertical scaled scores is unknown.

Due to the potential problems with scaled scores, this research has focused solely on using test scores in terms of raw percentage of questions answered correctly. We will address this issue further in Section 1.6; for now, we note that raw scores are the only possible method for analyzing the entire TAKS data set. Scaled scores stopped being reported in 2010, and vertical

scaled scores did not appear until 2009. By using raw scores, the benchmarks for passing and commended status will vary from year to year, but the variations are small enough (on the order of 5%) to be effectively discountable.

1.2.1 High Stakes Tests

Of all the TAKS tests, the ones administered in 5th, 8th, and 11th grade carry the most import. These grades are defined by the Texas legislature as *high-stakes* testing years, meaning students must pass the TAKS (now STAAR) tests in these grades in order to advance to the next grade [6]. Consequently, these grades in particular are the focus of many legislative policies that intend to help students raise their test scores (see Section 4.2). One such policy is that, if a student fails the TAKS exam when it is administered in April, he or she may retake the test up to two more times. If the student fails all three attempts, then he or she can be held back a year.[5]

The decision to retain students based on their standardized test scores is a controversial one. Proponents of the policy maintain that students who repeat a grade will have extra time to master the academic material, ensuring that they will not be overwhelmed once they advance to the next grade [53]. Those who oppose the policy claim that the psychological effects associated with being held back a year increase the chances that a student will experience

[5]There does exist a loophole where a student may proceed to the next grade if a committee unanimously determines that the student will likely perform at grade level by the end of the next school year, if given additional instruction [14].

developmental problems and/or drop out of school entirely. These opponents also state that at best retention has a minimal and transitory positive impact on score gains, and at worst retention actually has a negative impact on student achievement [53].

Research on the subject of retention has produced mixed results. Roderick and Nagaoka (2005) found that Chicago's students improved their test scores once its high-stakes testing program was implemented in 1996, but those students who were retained in the high-stakes years experienced short-term or negative achievement growth [40]. On the other hand, Peterson, DeGracie, and Ayabe (1987) reported that retained students in 1st through 3rd grade significantly improved their class standing for a period of up to two years past the retention year, although they feel that remediation is more effective than retention or promotion alone [37]. In fact, the findings of Lorence and Dworkin (2006) agree with Peterson et al. that retention can have a positive impact on the achievement score of students, if it is combined with additional instruction, assistance, and individualized educational plans [30].

While not directly testing the concept, the research in this dissertation agrees that extra instruction is more important than retention for improving the achievement of students. As we will discuss in Section 2.4.2 and Table 2.2, the retention rate across the state of Texas does not significantly increase as a result of high-stakes testing. However, we do observe an improvement in student test scores that persists for at least two years (see Section 4.3.2). It seems that at the very least, the threat of retention combined with remediation

led to a positive impact on student achievement.

1.3 Testing Theory and Reliability

In the classroom, teachers are mostly concerned with whether students can demonstrate that they have learned and retained the subject material. However, a student's performance on a test may be affected by factors that are independent of their knowledge of the material. For example, students may receive different scores on multiple-choice questions as opposed to essay questions; the teacher may accidentally misstate a question so that its meaning is unclear; or one version of a test may include leading questions that make it unintentionally easier than another version. This is true when discussing standardized tests as well, where it is expected that millions of students will take nearly identical versions of the same test that cover the same material and have the same difficulty level. The tests themselves cannot be the same from year to year (or cheating would be rampant) so they must be carefully constructed to be a true reflection of a student's knowledge.

Standardized tests are concerned with the concept of *reliability*, which assures that individuals' test scores remain relatively consistent if the test were repeatedly administered [23]. In other words, one goal of standardized tests is to make sure that variance among test scores is mostly attributable to the variance of the ability and knowledge of the students. However, any number of external factors (e.g. psychological, environmental) may affect a student's

test score. The *classical true score model* assumes that an observed test score X is a result of a true score component T (representing the expectation value of the student's score given an infinite number of tests) and a random error component E that includes mismarkings, lucky guesses, and other external factors [23]. In short, the classical true score model posits:

$$X = T + E$$

The *reliability index* ρ_{XT} of a test is a correlation coefficient that relates true and observed test scores. It is defined to be the ratio of the standard deviation of true scores to the standard deviation of observed scores [23]:

$$\rho_{XT} = \sigma_T / \sigma_X$$

σ_X is calculated from all possible observed scores that occur over many repeated testings, but true test scores cannot be observed without an infinite battery of tests, so σ_T is not directly measurable.

In comparison, the *reliability coefficient* $\rho_{X_1X_2}$ measures the correlation between observed scores on two parallel tests. Two tests are parallel if error variances are equal and all students have the same true score on both tests [23]. Since parallel tests ensure that the variances of observed test scores are equal (i.e. $\sigma_{X_1} = \sigma_{X_2}$), then the reliability coefficient is equal to the following

ratio:

$$\rho_{X_1X_2} = \sigma_T^2/\sigma_X^2$$

The reliability coefficient is equal to the proportion of observed score variance that can be attributed to true score variance. For example, if $\rho_{X_1X_2} = 0.91$ and the standard deviation of the observed score is 3 points, we could say the following:

- 91% of the observed score variance is due to true score variance
- The true score distribution has a standard deviation of $\sigma_T = \sqrt{0.91(9)} = 2.9$ points
- The correlation between observed scores and true scores is $\sqrt{0.91} = 0.95$

However, we cannot be certain that two tests are ever truly parallel, so the reliability coefficient as defined here is a theoretical concept much like the reliability index.

On the other hand, it is possible to estimate a lower bound for the reliability of a test using a value known as *Cronbach's alpha* or *coefficient alpha*, which is defined to be:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_X^2}\right)$$

where k is the number of items on the test, σ_i^2 is the variance of item i, and

Exam	Reliability
TAKS	.87-.90
TAKS-M	.82-.88
TELPAS reading (paper)	.93-.94
TELPAS reading (online)	.92-.95
Algebra I EOC	.92
Biology EOC	.91
Geometry EOC	.91

Table 1.3: Reliability estimates for TAKS exams

σ_X^2 is the total variance of the test [23].[6] The TEA uses a modified version of the coefficient alpha known as the *stratified coefficient alpha*. It is used when a mixture of item types appears on the same test [44] and it is defined to be:

$$\alpha' = 1 - \frac{1}{\sigma_X^2} \sum \sigma_{X_j}^2 (1 - \alpha_j)$$

where α_j is the coefficient alpha for each item type and $\sigma_{X_j}^2$ is the observed score variance for each item type [44]. Table 1.3 shows some reliability measurements for various exams as they have been estimated internally by the TEA (cf. Chapter 16, [44]). Here TELPAS refers to the Texas English Language Proficiency Assessment System, and TAKS-M is a modified version of the TAKS exam designed for students who are receiving special education services. The high reliabilities reported by the TEA for these tests indicate that they may be interpreted as valid assessments of actual learning.

[6] These variances may be determined from a single test administration, although the accuracy of the estimate will improve with an increasing number of responses.

1.4 Item Response Theory

Item response theory is central to the development of the TAKS test. It assumes that each student possesses a dimensionless *latent trait* θ that indicates the probability that a student will answer a question correctly. An *item characteristic curve* (or ICC) then maps the latent trait scores to a cumulative probability distribution [23]. For example, a latent score of 1 might map to 0.67 on a particular ICC, meaning that a randomly chosen student with latent score 1 has a probability of 0.67 of answering the question correctly. ICCs may take any shape, although some of the most common are S-curves or step functions.

Earlier research into item response theory used the *normal ogive* curve as the basis for its ICCs. The normal ogive curve is simply the cumulative distribution function of the normal distribution, where the latent trait values are equal to the distribution's z-score. The probability that a student with latent ability θ will answer item g correctly is given by:

$$P_g(\theta) = \int_{-\infty}^{a_g(\theta - b_g)} f(z)dz$$

where a_g and b_g are respectively a discrimination parameter and difficulty parameter for each item g [23]. Changing a_g affects the slope of the normal ogive curve, so it is directly related to the variance of the normal distribution; similarly b_g is equal to the midpoint of the curve (i.e. where half of the examinees would answer the item correctly) and is thus equivalent to the mean of the

normal distribution.

The normal ogive curve was eventually replaced by logistic models, where the probability of correctly answering a question is given by the logistic function:

$$P_g(\theta) = \frac{e^x}{1+e^x} \tag{1.1}$$

The value of x depends on the number of item parameters that are used to scale each ICC. In the two-parameter model, the parameters a_g and b_g are again used, and $x = Da_g(\theta - b_g)$ where D is an arbitrary constant [23]. If $D = 1.7$, then the two-parameter model is almost exactly the same as the normal ogive, and we have for Eq. 1.1 [29]:

$$P_g(\theta) = \frac{e^{Da_g(\theta-b_g)}}{1+e^{Da_g(\theta-b_g)}}$$

In the one-parameter logistic model, it is assumed that a_g is equal for all items; in other words, the slopes of all ICCs are the same. Then $x = Da(\theta - b_g)$, which allows for a rescaling such that $\theta^\star = Da\theta$ and $b_g^\star = Dab_g$. Eq. 1.1 then becomes:

$$P_g(\theta) = \frac{e^{\left(\theta^\star - b_g^\star\right)}}{1+e^{\left(\theta^\star - b_g^\star\right)}}$$

Georg Rasch used different concepts than the ICC to develop an equivalent model; hence, the one-parameter model is sometimes referred to as the *Rasch model* [51].

Finally the three-parameter model allows for correct responses due to

guessing. On multiple-choice questions, it may be assumed that there is a minimum percentage of students that will get the question correct regardless of latent ability, but the previous logistic models assume that $P_g(-\infty) \to 0$. The *pseudo-guessing parameter* c_g is introduced and Eq. 1.1 is modified to be [23]:

$$P_g(\theta) = c_g + \frac{(1 - c_g)\, e^{Da_g(\theta - b_g)}}{1 + e^{Da_g(\theta - b_g)}}$$

1.5 TAKS Scaling

To account for differences from one year's test to the next, the TEA uses the Rasch Partial-Credit Model (RPCM) to scale all raw test scores such that the panel-recommended Met Standard performance level is scaled to 2100 and the panel-recommended Commended Performance level is scaled to 2400. The RPCM defines the probability of a person n scoring x on an item i to be:

$$P_{xni} = \frac{\exp \sum_{j=0}^{x} (\theta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\theta_n - \delta_{ij})}$$

where m is the number of steps on item i, δ_{ij} are the step difficulties for item i, and θ_n is the student's proficiency level (cf. Chapter 15, [44]). However, the TAKS mathematics exams only have multiple-choice questions, so there are only two score categories (e.g. correct and incorrect) and the RPCM reduces

to the Rasch one-parameter Item-Response Theory model [38]:

$$P_{ni} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}$$

It should be noted that it is assumed here that the discrimination factor a_g is equal for all items. Items are field-tested using the entire state of Texas, garnering on the order of 100,000 responses per item (cf. Chapter 18, [44]). However, the process of determining the coefficients utilizes proprietary software and is not reported in the TEA's documentation.

The passing cutoff level for each test is determined using a modified item-mapping method (cf. Chapter 14, [44]). A group of panelists examined a booklet of proposed test questions that had been placed in order of increasing difficulty. Each panelist then placed a cutoff point at the question where a student who minimally Met the Standard should more likely than not be able to answer the question correctly [24]. (A similar cutoff point was used for the Commended level.) This process was repeated, and the panelists were shown their colleagues' analysis and responses to help inform their own decisions. After three rounds, the panelists made their final recommendations. Additional information about the methodological procedures is located in a 2002 report titled "Setting Standards on the TAKS Tests: A Modified Item Mapping Procedure", but this report is not publicly available through the TEA and any reported URLs are broken.

Once the cutoff scores for Met Standard and Commended had been

determined for each test, the Rasch ability scores were linearly transformed such that the Met Standard cutoff score scaled to 2100 and the Commended cutoff score scaled to 2400. The scale scores were developed as an alternative to the z-scores reported from the Rasch model because the new scale "...is easier to understand because it does not have negative numbers" (Chapter 15, [44]). The new scale ranges from approximately 1000 to 3200 for each test, and it includes weighted scores for the open-ended or essay questions on certain tests (e.g. exit level English language arts).

However, the scales are artificially adjusted for tests that include essay questions. A score of 2 or higher is required on the essay to achieve Met Standard on the writing and English language arts tests; a score of 3 or higher is required for Commended status. A student who fails to meet those requirements has their scaled score capped at one less than the appropriate level. So a student who receives a 0 or a 1 on the essay prompt cannot score higher than a 2099, regardless of his or her performance on the multiple-choice section; a student who receives a 2 on the essay cannot score higher than 2399 (cf. Chapter 15, [44]).

1.6 Raw vs. Scaled Scores

We choose to report test scores as raw percentages instead of scaled test scores in this research. There are several reasons for doing so. First, as mentioned in Section 1.2, TAKS does not use a single consistent scaling

process over its nine-year lifespan. Different measures are introduced and/or retired, and much of the scaling process is proprietary. In contrast, raw scores are usable in every year, and their meaning is transparent.

Second, the initial scaled scores used by the TEA were not intended to be directly comparable across tests. The TEA chose the fixed scale scores of 2100 for Met Standard and 2400 for Commended Performance so that students, parents, and the public could easily understand what individual scale scores might mean (Chapter 15, [44]). TAKS did not have a measure of student-level growth from grade to grade until the advent of vertical scaled scores in 2009. The Texas Growth Index was developed for the 2004-05 school year to estimate students' academic growth, but it was only a reliable statistical measure when aggregating students over campuses and/or districts (Chapter 12, [44]).

Third, we examined every conversion table from raw to scaled scores provided by the TEA. The number of questions required to pass the TAKS mathematics exam does not differ by more than two questions from year to year, and the overall number of test questions at each grade level remains constant. When the Met Standard score does vary by more than 5%, the tests tend to be alternate versions of the TAKS exam (e.g. online tests or alternate testing sessions). One exception is the 9th grade mathematics exam in 2010, where the Met Standard score differs by three questions from the year before.

Finally, as mentioned in Section 1.5, the TEA does employ artificial re-mapping of scores in cases where the student may score highly on the multiple-choice portion of a test but score poorly on the essay question. This may lead

to discontinuities in the data set which would break our methodology. Score re-mapping does not directly affect this research since it focuses solely on TAKS mathematics tests, which are entirely multiple choice. However, the STAAR standardized test uses open-ended griddable questions on its mathematics exam; this must be accounted for in future research.

1.7 Hierarchical Linear Modeling

The No Child Left Behind Act inadvertently created a windfall of data for educational researchers, but any statistical analysis of this data requires a robust framework. One commonly used tool is *hierarchical linear modeling* (also known generally in sociology as *linear multilevel modeling* [32]). This technique looks to exploit the hierarchical nature of the educational system, where students are grouped into classrooms, classrooms are grouped into schools, schools are grouped into districts, and so on. By increasing the "level" of the model, additional sets of regression coefficients are used to describe interactions between persons and persons, persons and schools, schools and districts, etc. Each level of the model is formally represented by its own submodel that specifies how its variables influence relations at other levels [18].

An example of a simple two-level hierarchical linear model examines the relationships between students and schools. At Level 1, the relationship between a student-level predictor (e.g. socioeconomic status) and a student-

level outcome (e.g. test score) can be given by:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}$$

where for a student i at a school j, X_{ij} is the predictor variable, Y_{ij} is the outcome, the β_{ij} are the fitting parameters, and r_{ij} is a random noise term [18]. If students are randomly assigned to schools, then this linear regression might lead to effective interpretations of school effects; however, this is rarely the case. Level 2 of this simple model would then account for school-level effects (e.g. public versus private schooling) by fitting the regression coefficients as:

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1} W_j + u_{kj}$$

Here the regression coefficients are γ_{kl}, the random noise term is u_{kj}, and the school-level effects are represented by W_j [18]. By arranging for the student-level regression coefficients to be dependent on school-level effects, the relative importance of each level's effects can be isolated.

Another example of how hierarchical linear modeling may be used is found in Miyazaki and Raudenbush [34]. In it, the authors utilize the National Youth Survey to determine how young persons' attitudes towards deviant behavior change as a function of age. The survey lasted for five years, but the age of the people interviewed ranged from 11 to 21. An initial suggestion might be to average together the answers of all people of a particular age, but

doing so could neglect historical effects that would cause one cohort to answer differently from another.[7] Hierarchical modeling is proposed here as a way to combine the data so that the entire age range is represented and any cohort effects, if they exist, will be immediately evident.

The model suggested by the authors is a two-level hierarchical model where the first level accounts for personal development as a function of age, and the second level accounts for development as a function of the cohort of the student. Because the graph of attitude versus age appears quadratic, they assume a model where individual change is given by:

$$Y_{ti} = \pi_{0i} + \pi_{1i}(a_{ti} - \bar{a}_i) + \pi_{2i}(a_{ti} - \bar{a}_i)^2 + \epsilon_{ti}$$

This is known as a *quadratic growth model* [18]. For person i and time t in this model, Y_{ti} represents the attitude score, a_{ti} is the age of the person, and $\overline{a}_i$ is the average age of the cohort to which person i belongs. In addition to being appropriately quadratic as a function of age, this Level 1 model includes fitting coefficients (represented by π) and a random error term ϵ_{ti} that is assumed to be normally distributed with mean zero.

Note that the model at this level assumes nothing about the cohort of each person. All the terms describe the age of the person, but any historical effects would not be observable. To account for this possibility, a second level

[7] While not covered in the scope of this paper, an example might be how students age 15 might behave pre- or post-Hurricane Katrina, or pre-/post-September 11th.

is added where all cohorts are compared to the first. Any significant deviation would suggest that cohort effects exist independently of age. The model is adjusted so that the fitting parameters are defined by the following:

$$\begin{aligned}
\pi_{0i} &= \beta_{00} + \sum_{j=1}^{6} \beta_{0j} d_{ji} + u_{0i} \\
\pi_{1i} &= \beta_{10} + \sum_{j=1}^{6} \beta_{1j} d_{ji} + u_{1i} \\
\pi_{2i} &= \beta_{20} + \sum_{j=1}^{6} \beta_{2j} d_{ji} + u_{2i}
\end{aligned}$$

In the second level, $d_{ji} = 1$ if person i belongs to Cohort $(j+1)$ where $j = 1, \ldots, 6$, and $d_{ji} = 0$ otherwise; this establishes that all persons in all cohorts are being compared to Cohort 1. Consequently all fitting parameters π are linear combinations of values of β and another random effect term given by u. This second level allows for the π coefficients to account for cohort effects while the first level models attitude as a function of age. This method of nesting different effects in different levels of the model is a key factor in hierarchical linear modeling.

While effective at estimating both fixed coefficients and random effects, there are three key problems with this method and how researchers often present its results. First, coefficients are typically assumed to be linear with respect to one another. In the example given here, there are no terms that allow

for possible interactions between more than two cohorts. This is a limitation of the model, but the severity of it depends on the system being analyzed.

A second limitation of hierarchical linear modeling is the assumptions that must be made in order to model a system. Miyazaki and Raudenbush made an assumption that the attitude towards deviant behavior behaved quadratically with age. However, one detail that has been left out is that Y_{ti} did not represent the score reported on this National Youth Survey for each person, but rather the natural logarithm of that score. They chose to do this to reduce skewness in the results and introduce normality, but it remains that the actual scores follow some type of exponential raised to polynomial time. While it is not necessarily incorrect to adjust or transform data for the purposes of modeling, any *a priori* assumptions must be taken with caution. Another example of this is the assumption of polynomial time; the authors concluded that a cubic term was statistically unnecessary, but no mention is given of testing other function forms.

Finally, interpreting the technical results of statistical studies may be a daunting task to those who have little or no expertise in mathematics. Table 1.4 gives an example of a result table from Miyazaki and Raudenbush. While they may be interpreted by statisticians and other experts in the field, politicians and administrators are usually the ones that need to understand these studies in order for any changes to be effected. Dense tables of coefficients, even when interpreted by experts, can lead to misunderstandings over the conclusions of those studies. In the worst case scenarios, politicians/ad-

Fixed Effect	Coefficient	SE	T	p
Level 1 intercept: π_0				
Intercept: β_{00}	0.325	0.0153	21.23	.000
d_2, β_{02}	0.067	0.0224	3.00	.003
d_3, β_{03}	0.119	0.0223	5.32	.000
Linear change: π_1				
Intercept: β_{10}	0.0648	0.0050	13.04	.000
d_2, β_{12}	-0.0158	0.0072	-2.20	.027
d_3, β_{13}	-0.0229	0.0068	-3.39	.001

Table 1.4: Reproduction of select results from Miyazaki and Raudenbush [34]

ministrators may actually mistrust the conclusions of researchers since they may not have the expertise to understand the results on their own. Ideally, the presenters would be clear in their presentations, and the audience would possess all the necessary tools to fully comprehend the data being set before them. This is far from what happens in reality.

Students (and people in general) are complex systems that are subject to many factors, variables, effects, and moods at any time. Essentially, populations are nonlinear systems. While a linear system might be made sufficiently complex by adding an indeterminate number of variables, it might be more efficient (and simpler) to avoid making assumptions about linearity in the first place. The model of Marder and Bansal is one such nonparametric model; that is, it makes minimal *a priori* assumptions about the distribution of student scores or how score changes respond to external variables.

Chapter 2

Methodology

2.1 Inspiration

Marder and Bansal created a model for analyzing longitudinal data [31]. Its inspiration is rooted in fluid mechanics, and consequently many terms from that science have been co-opted into this model (although they are not necessarily defined the same way).

Consider the following analogy: A straight horizontal pipe has water flowing through it. On a macroscopic scale, the water moves horizontally, but individual water molecules may not be moving horizontally at any given time. However, it cannot be said that the motion of a water molecule is random; the bias of motion is along the flow of the water through the pipe, and it is less likely for water molecules to start bouncing back and forth off the edges of the pipe. The position of the water molecules is time-dependent and semi-deterministic (i.e. any position is partially determined by where the molecule was at the prior moment).

This analogy can be extended to the test scores of students. First, test

scores are semi-deterministic; the score that a student receives in one year will be related to what that student scored in a prior year. If a student scores 70% on a standardized test in fourth grade, it is highly unlikely (though not impossible) that that student will score a 20% on the next test in fifth grade. It is much more likely that the student will score somewhere between 60% and 80% on the fifth grade test, with decreasing likeliness at scoring outside that range.

However, this is not a strictly first-order Markov process. A student may score 80% and 30% on consecutive tests, but it is likely that only one of the two scores reflects the student's understanding of the material being tested. Estimating the probability distribution of future scores does depend on past history, as well as variations in grade and score bin. The extent to which past history is important beyond a first-order approximation will be explored in Section 5.2.1.

For any student, there exists a probability distribution of what he or she will score in a given year, based on that student's scores in prior years. It is difficult if not impossible to construct a probability distribution for individual students, so this model aggregates students together based on demographic factors. To equate the analogies, the students are the "molecules", their test scores are their "positions", and the students' grades (i.e. fourth grade, fifth grade) are "time". The "pipe" in this model is the progress of students through the Texas educational system.[1]

[1] On a personal note, I get asked about my research quite a lot when people find out

2.2 Fokker-Planck Equation

While analogies are useful for understanding the basic idea, no model in physics would be acceptable without mathematical underpinnings. Here the basic analogy with fluid mechanics breaks down; basic equations from fluid mechanics like the Navier-Stokes equation or the Euler equations have no meaning here. An attempt to find a Reynolds number would be ill-defined. (It is worth mentioning that if the students' scores *were* to be treated as an actual fluid, it would definitely be compressible.) Instead, Marder and Bansal left fluid mechanics behind and used a more general dynamic tool.

The Fokker-Planck equation is used to describe the evolution of probability density functions. Its basic form in one dimension is given by:

$$\frac{\partial f}{\partial t} = -\frac{\partial}{\partial x}(\mu f) + \frac{\partial^2}{\partial x^2}(Df) \tag{2.1}$$

where μ is the *drift* of the function $f(x,t)$ and D is the *diffusion* of f. Both drift and diffusion may be position- and time-dependent.

We want to construct a Fokker-Planck equation for our system so that we can describe how the probability distribution of scores evolves as we look at different years, grades, and subsets of the population. Let s_t^α be the test score

that I am a graduate student. When I describe this model to them, their eyes inevitably light up regardless of their own studies/interests. It poses a very unique way of looking at longitudinal data and educational data in particular, and it is easy enough to intuitively understand that laypeople grasp the concepts almost instantly. The fact that this model is conceptually easy to understand on top of being mathematically rigorous has been a highlight of my research.

s of a student α in a given year t. Test scores may range from 0% to 100%, but it is more useful to bin the scores in some way. One reason for doing so is that since the number of test questions on each TAKS math test varies from grade to grade, it is possible to achieve a test score percentage in one grade and not in others. There is also legislation that prevents us from identifying students' scores exactly; we will discuss this later in Section 3.4. Let S_k be the k'th boundary of a score bin, where $S_k = \frac{k}{10}100\%$, $k \in [0, 1, \ldots, 10]$. So $S_2 = 20\%$, $S_3 = 30\%$, and a score s_t^α is defined to be in a score bin k when $S_k < s_t^\alpha \leq S_{k+1}$.[2]

This decile binning is useful for several reasons. First, each score bin now corresponds with a range of ten percentage points, which is very natural to discuss with people who are used to grades being assigned on a ten-point scale (90%-100% = A, 80%-89% = B, and so on). Second, while the TAKS cutoffs for passing and commended status vary from year to year, the passing cutoff consistently remains around 60%, and the commended cutoff consistently stays around 90%. The decile bins match these delineations nicely. This research has experimented with other types of binning (most notably five equally-sized bins) but none has been as convenient as the decile binning.

Let $A_{t,g,k}$ be the set of students who are in grade g in year t and who have a test score in bin $k \neq 0$, and let $N_{t,g,k}$ be the cardinality of $A_{t,g,k}$. Consider the students in $A_{t,g,k}$ who then advance to the next grade $g + 1$ in

[2]We do not use gain scores for two reasons: there is no concept of a pretest or posttest, and tests in different years focus on different concepts.

the year $t+1$, and upon taking the next year's test, their scores fall in the score bin $k' \neq 0$. For some of those students, their scores will be in the same bin (e.g. scored between 80% and 90% in both years t and $t+1$) and $k = k'$. The number of students that score in the same bin from one year to the next is surprisingly high; as mentioned earlier, scores vary but they rarely change by large degrees. Consequently, if one were to ask what $N_{t+1,g+1,k}$ is, a good starting estimate would be $N_{t,g,k}$.

We separate those students in score bin $k = 0$ from the rest of $A_{t,g,k}$ because it is very rare to see students who actually scored a zero on a given test. On a test of approximately 40 questions with four possible choices and no penalty for guessing, the chance of scoring a zero randomly is approximately 1 in 100,000. However, it is very common to find students whose scores are *reported* as zero. There are many reasons for doing so (see Section 3.2 and Table 3.2), but we choose to isolate those students from the rest of the data because their scores are artificial. To compensate, we include a *loss term* defined as $\triangle_{t,g,k} = R_{t,g,k\to 0} - R_{t,g,0\to k}$, where $R_{t,g,k\to 0}$ is defined to be the number of students who were removed from the data set between year t and year $t+1$, and $R_{t,g,0\to k}$ is defined to be the number of students who were added to the data set between those years.

The differences between $N_{t,g,k}$ and $N_{t+1,g+1,k}$ can be attributed solely to those students who had different scores between years t and $t+1$. Let $R_{t,g,k\to k'}$ be the number of students who have a test score in score bin k in year t and grade g, and who have a test score in score bin k' in year $t+1$ and

grade $g+1$. Then the difference between $N_{t,g,k}$ and $N_{t+1,g+1,k}$ can be expressed mathematically as:

$$N_{t+1,g+1,k} = N_{t,g,k} + \sum_{k'}(R_{t,g,k'\to k} - R_{t,g,k\to k'}) - \triangle_{t,g,k} \tag{2.2}$$

In other words, take all the students who scored k in year t, add all those that scored k in year $t+1$, and subtract all those that didn't score k in year $t+1$. Also, we subtract the loss term to represent the net number of students that disappear from the data set between years t and $t+1$. To turn this into a Fokker-Planck equation, the notation needs to be adjusted; instead of using k', other scores can be represented as a function of $\delta k = k' - k$. This notation allows Eq. 2.2 to be rewritten without loss of generality as:

$$N_{t+1,g+1,k} = N_{t,g,k} + \sum_{\delta k}(R_{t,g,k-\delta k\to k} - R_{t,g,k\to k+\delta k}) - \triangle_{t,g,k} \tag{2.3}$$

Note that the first k' is written as $k - \delta k$ and the second k' is written as $k + \delta k$. This may lead to ridiculous values for k' such as -30% or 120%, but those R values are simply zero because no students obtain those scores. Eq. 2.3 is still an exact equation describing this system.

Since scores do not typically vary by large degrees from one year to the

next, Marder and Bansal made the assumption that R is a slowly varying function of k [31]. This allows for a second-order Taylor expansion of $R_{t,g,k-\delta k\to k}$ to be written as:

$$\begin{aligned} R_{t,g,k-\delta k\to k} &\approx R_{t,g,k\to k+\delta k} - \delta k\frac{\partial}{\partial k}R_{t,g,k\to k+\delta k} \\ &\quad +\frac{1}{2}\delta k^2\frac{\partial^2}{\partial k^2}R_{t,g,k\to k+\delta k} \end{aligned} \tag{2.4}$$

Using this in Eq. 2.3, we get:

$$\begin{aligned} N_{t+1,g+1,k} &= -\triangle_{t,g,k} + N_{t,g,k} \\ &\quad +\sum_{\delta k}\left(-\delta k\frac{\partial}{\partial k}R_{t,g,k\to k+\delta k} + \frac{1}{2}\delta k^2\frac{\partial^2}{\partial k^2}R_{t,g,k\to k+\delta k}\right) \end{aligned} \tag{2.5}$$

$$\begin{aligned} \therefore N_{t+1,g+1,k} - N_{t,g,k} &= -\triangle_{t,g,k} - \frac{\partial}{\partial k}\left(\sum_{\delta k}\delta k\, R_{t,g,k\to k+\delta k}\right) \\ &\quad +\frac{\partial^2}{\partial k^2}\left(\frac{1}{2}\sum_{\delta k}\delta k^2 R_{t,g,k\to k+\delta k}\right) \end{aligned} \tag{2.6}$$

$$\therefore \frac{\partial N_{t,g,k}}{\partial t} = -\triangle_{t,g,k} - \frac{\partial}{\partial k}\left(v_{t,g,k}N_{t,g,k}\right) + \frac{\partial^2}{\partial k^2}\left(D_{t,g,k}N_{t,g,k}\right) \tag{2.7}$$

This looks very much like the one-dimensional Fokker-Planck equation from Eq. 2.1. One difference is that instead of representing drift as μ, this term

is referred to as *velocity* and it is denoted with a v. It is possible to go from Eq. 2.6 to Eq. 2.7 because δk is independent of k (and its derivatives). To finish the transformation, the velocity and diffusion terms need to be defined in the following manner:

$$\begin{aligned} v_{t,g,k} &= \sum_{\delta k} \delta k \, \frac{R_{t,g,k \to k+\delta k}}{N_{t,g,k}} \\ D_{t,g,k} &= \frac{1}{2} \sum_{\delta k} \delta k^2 \frac{R_{t,g,k \to k+\delta k}}{N_{t,g,k}} \end{aligned}$$

Velocity, therefore, is the sum of all score changes multiplied by the fraction of the total population that had those score changes. This notation was introduced in Marder and Bansal; an alternate way of writing these terms is by considering the individual test scores s_t^α in a given set $A_{t,g,k}$. In this notation, velocity and diffusion are given by:

$$\begin{aligned} v_{t,g,k} &= \frac{\sum_{\alpha \in A_{t,g,k}} (s_{t+1}^\alpha - s_t^\alpha)}{N_{t,g,k}} \\ D_{t,g,k} &= \frac{1}{2} \frac{\sum_{\alpha \in A_{t,g,k}} (s_{t+1}^\alpha - s_t^\alpha)^2}{N_{t,g,k}} \end{aligned}$$

It is worth noting that $N_{t,g,k}$ is an integer quantity, not a probability distribution. One might wonder if it can appropriately be called a Fokker-Planck equation in that case. An inherent assumption in this derivation is that the number of students is conserved over time. The definition of $A_{t,g,k}$ includes those students with nonzero scores in both years, but the loss term

includes everyone else. Therefore, the number of students is conserved over time because every student is either mapped to an R term or the loss term. Let $N = \sum_k N_{t,g,k}$ be the total number of students in year t and grade g such that all students in the loss term are represented by the score bin $k = 0$. Then $N_{t,g,k}/N$ is a probability density as a function of k, and we may divide both sides of Eq. 2.7 by N to turn it into a true Fokker-Planck equation.

2.3 Dirac Formalism

The terms in Section 2.2 are used to illustrate the mathematical formalism behind the theory. In practice, much of the analysis is conducted using multi-dimensional tensors where the axes correspond to year, grade, score bin, or other demographic variables. The use of these tensors eventually led to an insight that Dirac notation might be a useful alternative for describing our system.

Let $|S\rangle = \sum_i |S_i\rangle$ be the orthonormal basis of our vector space so that $\langle S_i|S_j\rangle = \delta_{ij}$. Each $|S_i\rangle$ corresponds to S_k and is usually a percentile rank. We can then define an operator $\hat{S} = \sum_i S_i \, |S_i\rangle \, \langle S_i|$ such that:

$$\begin{aligned} \hat{S} \, |S_i\rangle &= S_i \, |S_i\rangle \\ \langle S_j| \, \hat{S} &= \langle S_j| \, S_j \\ \langle S_j| \, \hat{S} \, |S\rangle &= S_j \end{aligned}$$

Symbol	***Meaning***
t	An integer denoting the year in which a test is taken. When a test is taken in an academic year such as 2009-2010, we use $t = 2010$.
s_t	A test score in year t, in units of percentage of maximum score.
s_t^α	The test score of student α (an integer) in year t in units of percentage of maximum score. When students take multiple administrations of the exam during the year, we choose the maximum.
g_t^α	The grade level of student α in year t.
S_k	The k'th boundary of bins used to make scores discrete: $S_k = \frac{k}{10}100\%$, $k \in [0, 1, \ldots 10]$. A score s_t is in bin k when $S_k < s_t \leq S_{k+1}$.
$A_{t,g,k}$	A set of students who in year t are in grade g, whose test score is in bin $k \neq 0$, who advance to grade $g+1$ the following year, and who have nonzero score the following year.
$N_{t,g,k}$	The cardinality of the set $A_{t,g,k}$ (i.e. the number of students in year t, grade g, and bin k).
$v_{t,g,k}$	The average score change of students in year t, grade g, and bin k (in set $A_{t,g,k}$).
$\overline{s}_{k_0,g_0,t_0\to t}$	The average score in year t of students who in year t_0 had score given by k_0 and were in grade g_0.
$\mathsf{S}_{t'}^{s,t}$	The score in year t' of a trajectory passing through score s in year t.

Table 2.1: Table of notations and conventions used. [15]

As before, $|S_0\rangle$ corresponds to the scores between 0% and 10%, $|S_1\rangle$ corresponds to the scores between 10% and 20%, and so on.

Let us take a sample case where a population of students has some distribution of scores in year A and some other distribution of scores in year B. Define N_i^A to be the number of students in year A with a score corresponding to S_i. We may then define a vector $\left\langle N^A \right| = \sum_i N_i^A \langle S_i|$ that represents the distribution of student scores in year A with respect to our basis.

Let r_{ij} represent the probability that a student with score S_i in year A has a score S_j in year B. We know the following:

$$\begin{aligned} \sum_j r_{ij} &= 1 \\ \sum_i r_{ij} N_i^A &= N_j^B \end{aligned}$$

Let $\hat{r} = \sum_i \sum_j r_{ij} |S_i\rangle \langle S_j|$. Take note that $\hat{r}$ is a right stochastic matrix. Then $\left\langle N^A \right| \hat{r} = \left\langle N^B \right|$.

Proof:

$$\begin{aligned}
\left\langle N^A \right| \hat{r} &= \sum_k \sum_i \sum_j N_k^A r_{ij} \left\langle S_k | S_i \right\rangle \left\langle S_j \right| \\
&= \sum_i \sum_j N_i^A r_{ij} \left\langle S_j \right| \\
&= \sum_j \left(\sum_i N_i^A r_{ij} \right) \left\langle S_j \right| \\
&= \sum_j N_j^B \left\langle S_j \right| \\
&= \left\langle N^B \right|
\end{aligned}$$

Define $R_{ij} = N_i^A r_{ij}$ to be the number of students that move from score S_i in year A to score S_j in year B. We consequently define an operator $\hat{R} = \sum_i \sum_j R_{ij} \left| S_i \right\rangle \left\langle S_j \right|$ with the following properties:

$$\begin{aligned}
\hat{R} \left| S \right\rangle &= \left| N^A \right\rangle \\
\left\langle S \right| \hat{R} &= \left\langle N^B \right| \\
\left\langle S_i \right| \hat{R} \left| S_j \right\rangle &= R_{ij}
\end{aligned}$$

$$\sum_j R_{ij} = \sum_j N_i^A r_{ij} = N_i^A$$

This alternate formalism may lead to other non-traditional ways of

analyzing our data set. We will touch briefly on some of those potential applications in Section 5.2.4.

2.4 Visualizations

This model can be utilized for statistical analysis of longitudinal data sets. As mentioned in Section 1.7, a common method for analyzing educational effects is through hierarchical linear modeling, but this method presents problems both in terms of assumptions of linearity and representations to laypeople. We have developed visualizations that solve both of these problems. They are based on the model of Marder and Bansal, and are inherently nonlinear and nonparametric. The results can be presented in pictorial form. While some explanation is still needed to interpret the graphs, visualizations can be more accessible and help others understand the analysis being presented.

2.4.1 Snapshot Flow Plots

The first type of visualization was introduced in Marder and Bansal [31], and referred to as a *snapshot* flow plot. Flow plots use arrows to show how students' test scores change on average from one year to the next. More specifically, snapshot flow plots show a given system over a period of two years.

Refer to Figure 2.1 for an example of a snapshot flow plot. This shows the change in test scores for all Hispanic students in the state of Texas from 2009 to 2010. It is very common in these flow plots to separate students by

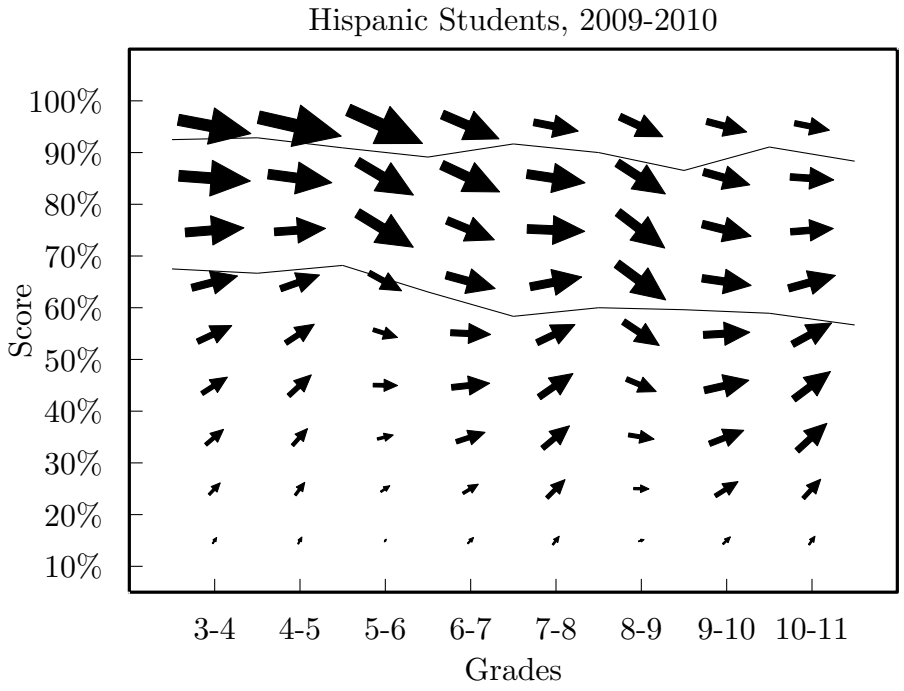

Figure 2.1: Example of a snapshot flow plot

some demographic variable, whether that be ethnicity, sex, economic status, or campus. The horizontal position of each arrow corresponds to the grade g of the students in the year t (here, $t = 2009$). The vertical position of each arrow corresponds to a score bin k; all students represented by a given arrow are placed in the same score bin k for the year t. The area of each arrow is proportional to the number of students $N_{t,g,k}$ in each score bin. Finally, the angle of each arrow is equal to the arctangent of the velocity $v_{t,g,k}$. In other words, the vertical height of each arrow is proportional to the average score change of the students in that bin from t to $t + 1$; each arrow points at the average score of its students in the year $t + 1$. If each arrow were enlarged so that its horizontal component were one unit of t in length, then the vertical component of each arrow would be exactly $v_{t,g,k}$. To aid viewers in assessing how students are performing with respect to the TAKS test, two thin black lines have been added to show the percentile scores corresponding to the tests' passing and commended score cutoffs for each grade. Figure 2.1 is only one example of a snapshot flow plot, but regardless of the year, most passing cutoff scores track near 60% and most commended cutoff scores stay near 90%.

Interpretation of these graphs is fairly simple, after a little instruction. Each column of arrows represents a particular grade, and the respective areas of the arrows shows its score distribution. Note that most arrows at the bottom of the graph are very small when compared to the arrows at the top of the graph, showing that many more students do well on the TAKS test than score abysmally. One can compare how students do between grades by comparing

one column to another; the distribution of scores for the 5th to 6th grade transition is much more top-heavy (top arrows are much larger than bottom arrows) than the relatively uniform distribution for the 10th to 11th grade transition. The angles of the arrows shows how the test scores changed from 2009 to 2010.

All arrows at the top of the graph point down, and this is to be expected; for students that score between 90% and 100%, scores will decrease on average in the next year. This is for two reasons. First, the peaks of the statewide score distributions tend to lie between 50% and 70%. Some students may have a true knowledge state closer to the average, but they are in the top score bin due to factors outside of the test's control (e.g. guessing, an exceptional teacher for a year; see Section 1.3). On average, these students will tend to regress to the mean of the score distribution, causing the top arrows to point downward. Second, since a student cannot score above 100%, anyone who gets a perfect score on one test has to score either the same or lower in the next year. When examining tens of thousands of students, the net effect is always a decrease in average score.

As another example, compare the angles of the arrows from the 3rd to 4th grade transition to the 5th to 6th grade transition. The 3rd to 4th grade arrows either point down at a very shallow angle, or they point up. This indicates that students in 4th grade do as well or better when compared to their 3rd grade scores. In contrast, the large arrows in the 5th to 6th grade transition (where most of the students are binned) point sharply downward.

The transition from elementary school to middle school is not an easy one for many students, and their test scores drop precipitously in 6th grade. (Some schools treat 6th grade as a part of elementary school instead of middle school; for the 2010-11 school year, there were 171 schools that taught kindergarten to 6th grade and 549 schools that only teach kindergarten through 5th grade [10]. We will explore this topic briefly in Section 5.2.2.) A similar effect occurs between 8th grade and 9th grade, when students transition into high school. In this case, the sharp declines are exaggerated due to a program known as the Student Success Initiative (see Section 4.2) but the effect is still present in all visualizations, regardless of whether the Student Success Initiative was active or not.

These graphs are used for qualitative analysis, not quantitative. Should someone wish to know the exact numbers behind any of the graphs or arrows, they may be quickly produced from the data. For most cases, flow plots can be used to give an overall picture of how a subset of students are performing on standardized tests. It is possible for one to integrate over the vector field produced by a flow plot and include diffusive terms to form streaklines. However, a person's eyes are fairly good at following the size of the arrows and the direction they point to visualize the overall pattern of student test scores. This is best seen when comparing two disparate groups of students and seeing how their flow plots change. Figure 2.2 is one such example of this. These flow plots compare economically disadvantaged students (defined as those receiving free lunches, reduced-price lunches, or some other form of financial

aid) against those that are financially well-off. Both subsets of students do well in the early grades, with a high concentration of students in the upper percentiles and very little downward flow. However, as students transition into middle school and experience the inevitable drop in test scores, economically disadvantaged students suffer a much larger drop than well-off students, and do not recover. Consequently, by 11th grade the economically well-off students are consistently outperforming the economically disadvantaged students.

2.4.2 Cohort Flow Plots

Snapshot flow plots are useful for giving an overview of a subset of students over a given two-year time period. However, it is important to note that many different grades of students are represented in the same plot. In Figure 2.1, the entire flow plot shows how Hispanic students performed between 2009 and 2010, but the columns focus on Hispanic 3rd graders, 4th graders, 5th graders, etc., from the same calendar years. If one wanted to observe how an event (for example, Hurricane Ike) affected a group of students as they progressed from 3rd to 11th grade, snapshot flow plots would be incapable of identifying any perturbations in the system.

Cohort flow plots are another method of visualizing the data that is more useful for answering questions like this. The basic construction of the cohort flow plot is identical to that of the snapshot flow plot; the area of the arrows is proportional to $N_{t,g,k}$, the vertical height of each arrow is proportional to $v_{t,g,k}$, and the vertical position of each arrow corresponds to the score bin k.

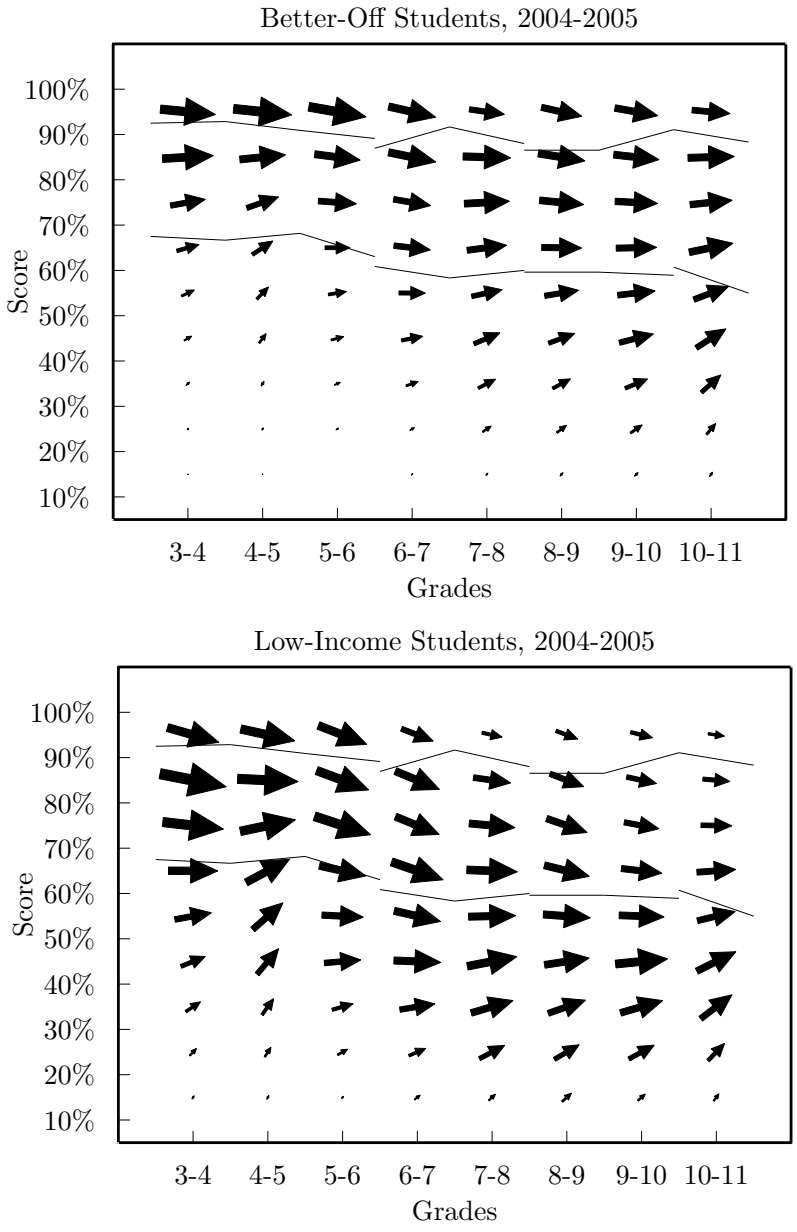

Figure 2.2: Snapshot flow plots comparing students who are economically disadvantaged and economically well-off. The discontinuities in the passing and commended cutoff lines is due to changes between 2004 and 2005 in the number of questions required to achieve passing or commended status on those tests.

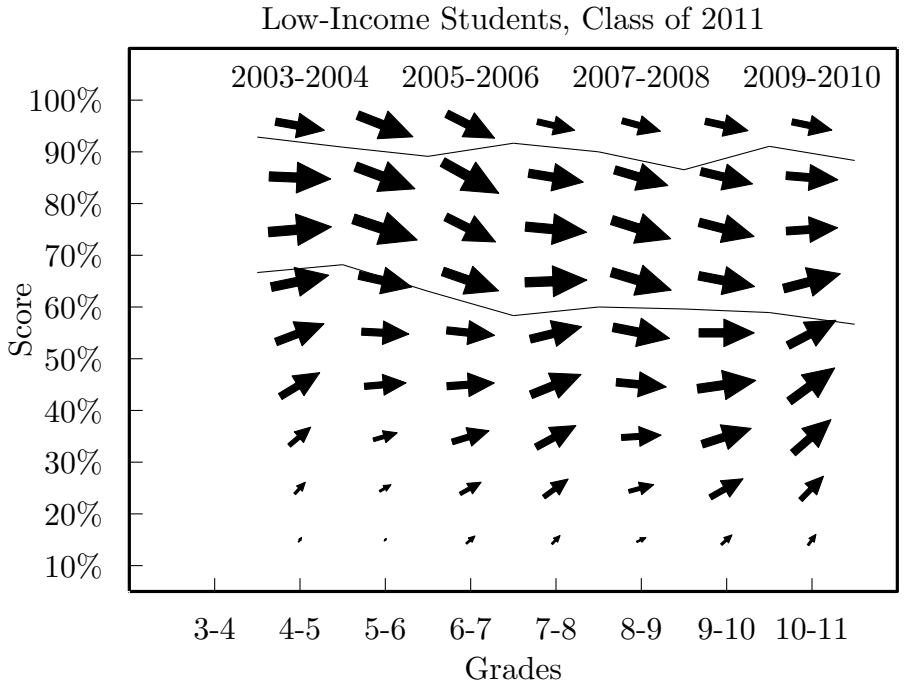

Figure 2.3: Example of a cohort flow plot

What changes for a cohort flow plot is that each column represents a different set of calendar years. As the grade transitions increment along the horizontal axis, the years increment as well. This means that each column represents the same set of students, at different points in their educational careers. This is shown by having the title reference their graduation year. Figure 2.3 is an example of a cohort flow plot; calendar years are included at the top of the graph to help illustrate how the years correspond to the grades/columns.

Visual integration of the arrows in a cohort flow plot does give an accurate representation of how that cohort's educational career changed over time. There are, however, a few disadvantages to using cohort flow plots over

snapshot plots. First, cohort flow plots take many years to develop. Cohorts in Texas take nine years to fully progress through the battery of standardized tests, but snapshot flow plots can be completed after only two years of data gathering. Texas has changed its version of its standardized test twice in the past thirteen years (see Section 1.2). Only the Class of 2012 will have taken the TAKS test every year from 3rd to 11th grade. Cohort flow plots may be impossible to fully develop for a cohort, depending on the number of grades one wishes to observe and the stability/volatility of the standardized test that is being examined.

Second, an implicit assumption of a cohort flow plot is that each column represents the same subset of students. However, this may not be the case. Students may drop out of school, or move to Texas from another state. Students may be held back a grade, or skip ahead one. In some cases, the data may be misrecorded such that the student is identified with the incorrect cohort.

In the case of retention, we need to determine whether, when looking at statewide populations, the number of students who are retained a grade is small when compared to the overall population. We can calculate the retention rate by seeing which students have the same grade in consecutive years. Table 2.2 shows those grades and years which had a retention rate greater than 4%. It must be noted that only the low-income students are listed in the table; when looking solely at economically better-off students, no year/grade combination had a retention rate higher than 4%. 9th grade is especially problematic as

it sees retention rates near 10% in every year. Those students' scores are still counted in each 9th-10th grade arrow of a cohort plot, but they are not present in the 10th-11th grade arrows. This is a sink for our system, and was touched upon briefly in Section 2.2 when discussing the loss term of the Fokker-Planck equation.

In the case of appearances and disappearances, the number of students who are permanently added or removed from a given cohort is small compared to the number of students whose test scores are marked zero for legitimate reasons. This will be expounded upon in much greater detail in Section 3.2, but the essence of the argument is that there are ways of identifying tests that were not scored for valid reasons, and we do not count those students in our plots because their zero scores artificially lower the average score changes across all bins. By eliminating those zero scores, we choose to focus on the large majority of students who received test scores in consecutive years.[3]

When dealing with large populations, the number of students who fall out of a given cohort is miniscule compared with the total population. However, if one wishes to apply exceedingly fine disaggregations to the model, the columns may not represent the same subset of students at all. Specifically, campus-level graphs may contain so few students that no value can be gleaned from a cohort flow plot, especially in the event of catastrophic perturbations

[3]Excluding students who had valid zero scores, the number of appearances and/or disappearances amounts to less than 3% of the population in every case, with one exception. For low-income students in 2009 in 9th grade, approximately 6% of the students had no score in the prior year.

Years of transition	Grade Level	Percent retained
2003-04	9th	11.2%
2003-04	10th	5.2%
2004-05	9th	11.9%
2004-05	10th	4.9%
2005-06	3rd	4.4%
2005-06	5th	4.8%
2005-06	9th	11.3%
2005-06	10th	4.8%
2006-07	9th	11.0%
2006-07	10th	4.6%
2007-08	9th	9.5%
2008-09	9th	8.4%
2009-10	9th	6.8%
2010-11	9th	8.9%

Table 2.2: Percentage of students held back in terms of grade level and year. Only those year/grade combinations with greater than 4% retention are noted here. These retention rates solely refer to students receiving free or reduced-price lunches.

to the system (e.g. Hurricane Katrina).

2.4.3 Streamline Plots

People seem to be remarkably adept at visually integrating a flow plot to get an idea of how the population changed over time. Of course, one could just numerically integrate over the arrows and present the results in graphical form. These plots are called *streamline* plots, where the name derives from using the exact same process that one would use to obtain particle streamlines from a velocity vector field. Linear interpolation across any grade level's $v_{t,g,k}$ values leads to a series of continuous functions $v_{t,g}(x)$, where $x \in [0, 1]$. The

boundaries are set by linear extrapolation, with a caveat that $v_{t,g}(0)$ or $v_{t,g}(1)$ cannot lead to impossible score values. If either $v_{t,g}(1) > 0$ or $v_{t,g}(0) < 0$, those functions are set to zero at the boundaries instead. This has not occurred for any set of $v_{t,g,k}$ that has been recorded in this research, but this check is still in place.[4]

Once the continuous $v_{t,g}(x)$ have been created, a starting score in third grade is chosen, and the appropriate $v_{t,g,k}$ is used to calculate those students' average score in fourth grade. From there, the process is repeated with the continuous $v_{t,g}(x)$ to get an estimate of the average score in fifth grade, and in all other grades. Following the lead of snapshot and cohort flow plots, the width of the streamlines is proportional to $\sqrt{N_{t,g,k}}$, so the area of each segment of the streamline corresponds to the population $N_{t,g,k}$.

Figure 2.4 shows an example of a typical streamline plot. The streamlines tend to grow very close together by 11th grade, but this effect is due to regression to the mean, and underestimates the actual differences between the top and bottom streamlines. This is a visualization problem similar to the one faced by snapshot and cohort flow plots; streamline plots are good for showing qualitative differences between populations, and if generated from snapshot flow plots, they can be assembled with only two years of data. However, the

[4] When first creating streamline plots, the boundary conditions of $v_{t,g}(0)$ and $v_{t,g}(1)$ were erroneously set to zero automatically. This was a reflex due to working with streamlines and assuming a no-slip condition as in fluid mechanics, but the educational system does not follow all the rules of a physical pipeline. Assuming that $v_{t,g}(0) = v_{t,g}(1) = 0$ leads to nonsensical interpolations near the boundaries for $v_{t,g}(x)$. It is worth noting just where and when adhering to the analogy can cause quite a few problems!

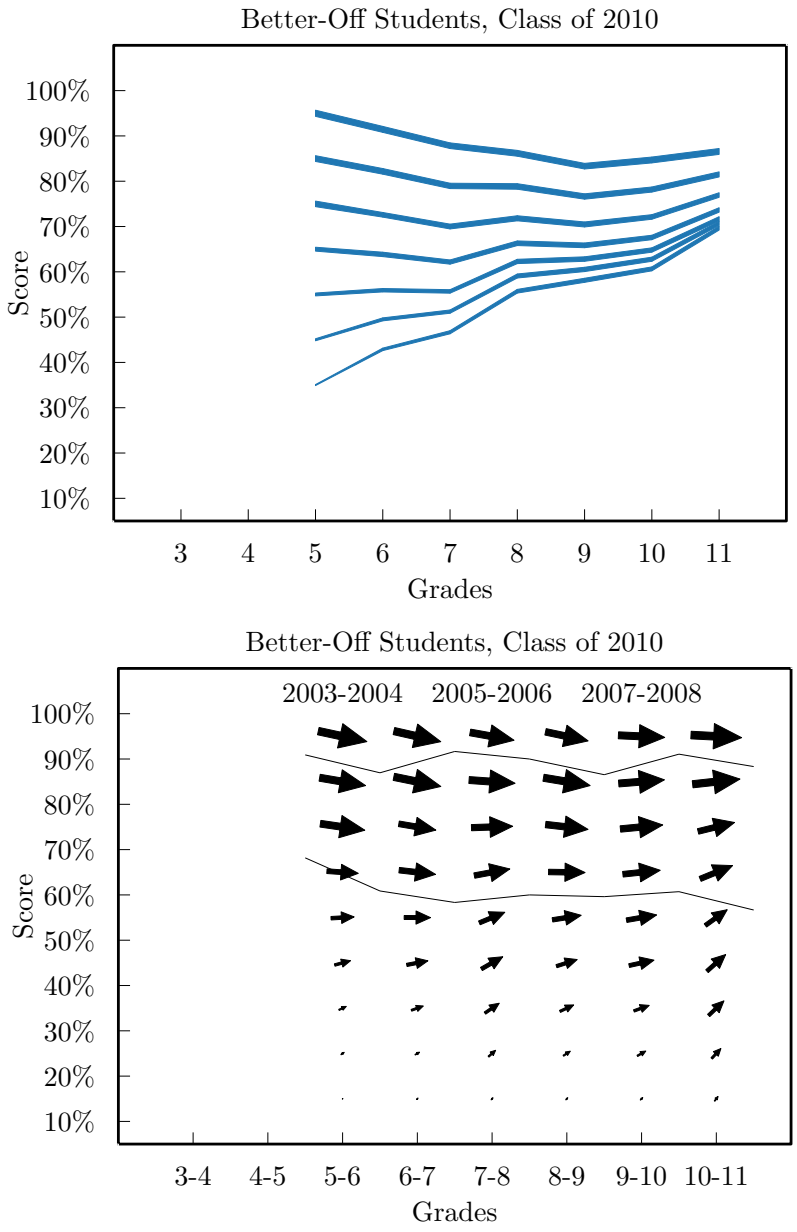

Figure 2.4: Example of a streamline plot and the cohort graph that generates it

results cannot always be taken at face value. Furthermore, streamline plots use linear interpolation and only examine score changes from year to year. Score changes that rely on two or more years of history are not considered here (see Section 5.2.1).

2.4.4 Trajectory Plots

Trajectory plots take their name from their counterpart in fluid dynamics. Like with streamline plots, the method of obtaining trajectory plots is identical to how one would trace the trajectory of a particle in a fluid flow; namely, just follow each particle over each timestep and see where it ends up. Formally, a trajectory plot begins by fixing a subset of students A_{t_0,g_0,k_0} at some initial year t_0, initial grade g_0, and initial score bin k_0. At all years $t > t_0$, the scores and grades are recorded. Trajectories plot the average score $\overline{s}_{k_0,g_0,t_0\to t}$ of the chosen subset of students for all years $t > t_0$. The width of the lines once again is proportional to $\sqrt{N_{k_0,g_0,t_0\to t}}$. It is not unusual for the width of the lines to vary slightly because not all students from A_{t_0,g_0,k_0} will have a record in every grade. Students may drop out, leave the state of Texas, or in many cases be subject to data recording errors.

Figure 2.5 shows an example of a trajectory plot; compare it to Figure 2.4 to see how streamline plots differ from trajectory plots. While they share the same qualitative properties, trajectory plots are much more accurate for showing how a subset of students A_{t_0,g_0,k_0} is performing on average at a later date $t > t_0$. They show exactly how students at an initial grade and perfor-

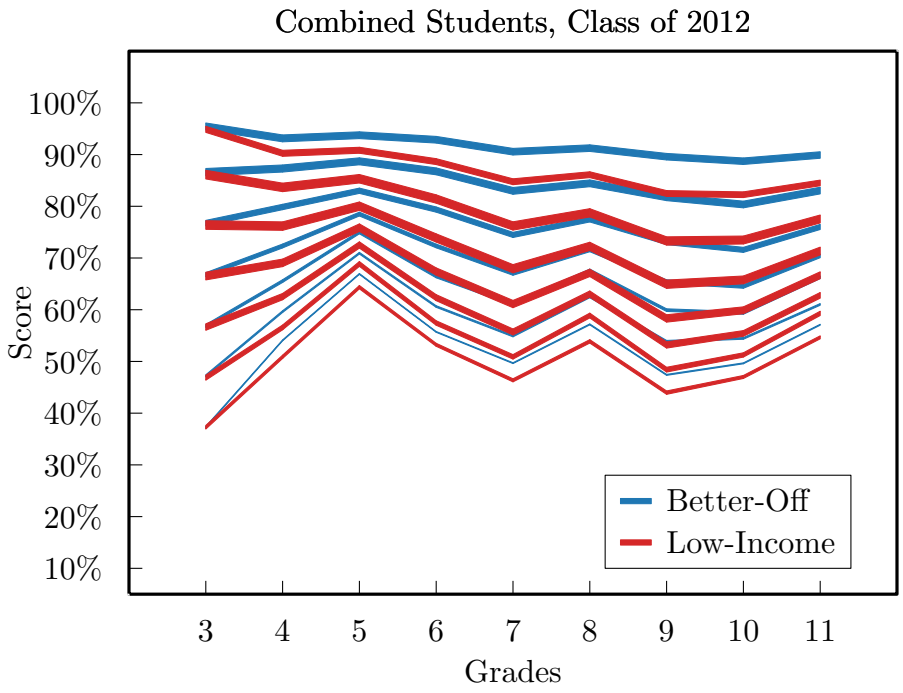

Figure 2.5: Example of a trajectory plot

mance level are testing by the end of their educational careers. However, much like cohort plots, trajectory plots take many years to generate. Politicians and administrators can rarely afford to allow nine years to elapse in order to make perfectly informed decisions about legislation and policy. The task of obtaining the most accurate possible trajectories from the smallest number of years of data is the subject of ongoing research.

Chapter 3

Data Analysis

3.1 Introduction

For computational physicists, an ideal world would be one in which data is easily collected, easily managed, and easily analyzed to produce results. Clearly, reality often refuses to adhere to these wishes. The problems with data collection/usage magnify when dealing with nonideal physical systems, scientific systems too messy for simplification, and with social systems that have little standardization. The difficulties presented by these systems are not necessarily impossible to overcome, but they do take significant time and energy to do so.

3.2 TAKS Data Set

The Texas Education Agency (TEA) has created contracts with Educational Research Centers (ERC) in the state of Texas. These ERC facilities are given special access to data that would normally be locked away under the Family Educational Rights and Privacy Act (FERPA). The Texas Schools

Project (TSP) at the University of Texas at Dallas (UTD) is one such ERC facility. Their servers contain a host of educational data from the state of Texas, including but not limited to the testing results from the TAAS, TAKS, and STAAR tests (for additional reference, see Section 1.2). Our research focuses solely on the TAKS data.

In its rawest form, the TAKS data is separated into individual files corresponding to every version of every test from 2003 to 2011. This includes separate files for variants of tests including online testing, Braille testing, and Spanish-language versions of the test. These files are initially saved in SAS format (a statistical programming language that is ubiquitous in the educational research community). The staff at the TSP was fortunately kind enough to place the files into a common CSV format, and once there the total length of all the files together is over 38 million rows. Each row corresponds to a single test record for a given student, where that student is marked with an anonymous ID number. In some cases, the student takes multiple versions of the same test, and multiple rows appear for that single student for that particular grade and year. In other cases, the student took a single test, but the subjects have been spread out across multiple rows (e.g. the scores for Reading are one line, but the scores for Mathematics and Science are on another line). The TAKS data set is also prone to inconsistencies from one year's file to the next. Some may be due to changes in the student's status (e.g. the student received reduced-price lunches in 2005 and 2006 but not 2007), others might be due to inaccurate demographic questions (e.g. biracial students se-

lect "Black/African-American" in 2008 and "White" in 2009), and still others are likely due to inaccurate data recording (e.g. a student supposedly takes the 8th, 9th, 3rd, and 11th grade tests in consecutive years).

These data files must be collected, formatted, and cleaned of as many errors as possible. The data ideally should be organized such that every row of the file contains the complete educational history for a single student. This is not a simple task, given that each student's records are spread across multiple CSV files with no way to quickly search for common IDs. Furthermore, the test format stayed more or less consistent from 2003 to 2007, but in 2008 the column headings changed for some of the categories (e.g. a student's individual responses on a math test changed from "M_IRSP" to "M_IRS"). In 2011, the TAKS test changed how it records race/ethnicity information (see Table 3.1); this has not affected our current research because there are no transitions from 2011 to 2012 to speak of, but it may become important if the results of this research will ever be used to link TAKS to STAAR.

I have had to perform this compiling and cleaning operation three times during our research, mostly due to updates as new data became available. Each version has required significant rewrites to the code to adapt to the changes in the TAKS data file format. During the second of these three occasions, I attempted to build the most complete longitudinal record for all the students that I could find. Specifically, I gathered information about all the various subject tests and included any column that occurred in any of the TAKS files. This bloated the total data set to almost 60 GB of plaintext CSV file

Pre-2011	2011 and later
1 = American Indian or Alaskan Native 2 = Asian or Pacific Islander 3 = African American 4 = Hispanic 5 = White, not of Hispanic Origin	I = American Indian or Alaskan Native A = Asian B = Black or African American H = Hispanic/Latino W = White P = Native Hawaiian or Other Pacific Islander T = Two or More Races N = No Information Provided

Table 3.1: Comparison of race/ethnicity options on TAKS test [46]

(much of it blank) and the operation took over 86 hours to run. In May 2013, I extensively rewrote much of our code to focus on memory efficiency, input/output optimizing, and culling unnecessary data. Specifically, I only selected columns from the TAKS data files that were relevant to our current research questions, such as how students' performances on the TAKS math test were affected by SSI (see Section 4.2). This reduced the preparation time to a scant 40 minutes. Figure 3.1 shows pseudo-code that illustrates the current process of putting the ERC's data into a manageable form.

Those reading this dissertation may wonder at our choice to use plain-text CSV files instead of more efficient database managing systems like MySQL. Initially, much of the code was written in a hybridized form of MySQL and Python. A publicly available version of the TAKS data set had been acquired by Dr. Michael Marder and stored in an SQL database on our local system.[1]

[1]This publicly available data set contained information from 2003 through 2007, but over

```
fields = ['ID2',
          'YEAR',
          'GRADE',
          'DISADV',
          ...]
file_paths = find_all_TAKS_files()
for file in file_paths:
    for row in file:
        info = grab_relevant_fields(row)
        select_temp_output(id_number)
        temp_output.write(info)
sort_all_temp_output_files()
sorted_data = combine_temp_files()
for row, next_row in sorted_data:
    if row[id] == next_row[id] \
    and row[year] == next_row[year]:
        merge_rows(row, next_row)
```

Figure 3.1: Pseudo-code showing the steps of converting ERC-formatted TAKS data to a CSV file suitable for our research

However, when I began working with the Texas Schools Project, I discovered that MySQL was not available on their servers for clients to use, and a working version was around six months away (and may have never been installed; see Section 3.4 for a description of the technical limitations encountered on this project). The difficulties in working with Texas's data set made us wonder how many more problems we would encounter if we tried to extend our model to other states' data. We concluded that it would be best to write our code to utilize CSV files, which are simple and almost universally adaptable on all systems. We traded database efficiency and faster real-time searches for flexibility and simplicity. This has had the added benefit of requiring new researchers to learn only a single programming language instead of two or more.

The step at which we reconcile errors in the data set is one of the largest assumptions we make in this model. In order to get our longitudinal data set in its desired format, we can only have one test score for a student in a given year. Most of the time, this poses no problem for our data. However, certain years do allow for retakes. Specifically, students in 5th, 8th, and 11th grades often have multiple test records due to the high-stakes nature of those tests (see Section 1.2.1). To reconcile these test scores together, we choose a student's score to be his or her best score in a given school year. This has the effect of artificially inflating test scores in high-stakes years; however, doing so allowed us to discover the effect of a Texas educational policy that was a perturbation to the system (see Section 4.2). Rewriting our code and data

30% of the rows had been replaced with null values due to FERPA restrictions.

formats to accommodate multiple tests in a year may be the focus of future research on this subject.

Initial attempts to build up a complete longitudinal record for each student resulted in an alarming number of values in our lowest score bin (where scores are between 0% and 10%). We discovered that we had made the assumption that any null values would be counted as a zero score, but the number of students scoring this low was incredibly large compared to the next lowest bin. It had the effect of reporting velocities far below what they should be and depressing the arrows on all our plots. Fortunately, further analysis of the TAKS data showed that there were common qualities to tests with no score. Specifically, each TAKS test included a "score code" that can be used to denote special circumstances surrounding the student or the testing environment. It turned out that tests with seven specific score codes were responsible for over 88% of all reported null values (see Table 3.2). Furthermore, over 97% of all tests with these score codes returned a null score. By excluding all of these score codes from our analysis, our vector fields became much less discontinuous. In the future, it is possible that score codes will be changed to reflect new populations of null scores, so Table 3.2 should be checked repeatedly.

3.3 Research Data Formats

The combined records of all the TAKS data files together amount to over 38 million rows. Reconciling multiple tests in a single year reduces it to

A	Absent
D	No information available for this subject
G	TAKS-Alt record
P	Previously Met Standard (Grades 3 and 5 and exit level retest administrations)
Q	Student did not take the TAKS mathematics test, do not score (Grades 3 and 5 February and Grades 4, 6, 7, and 8 April) (SDAA II)
X	Student is ARD exempt, do not score (exit level)
(no value)	No score code recorded

Table 3.2: Score codes responsible for the majority of null scores on TAKS tests

over 27 million rows, which correspond to 6.5 million unique students over the 2003-2011 time period. Once we have constructed our longitudinal data set, we need to extract useful information from it.

One of the most unique aspects of our research involves the data structures that we extract from the TSP. Instead of using spreadsheets, we use the Numpy package to create multidimensional tensors that contain years of information. Numpy allows us to have established data types that are linked together either as a dictionary (e.g. the dictionary keys are campus numbers) or just as another dimension of the tensor (e.g. an axis where the indices correspond to the ethnicity codes in Table 3.1).

Velocity grids are one of the main tools we use for creating snapshot flow plots, cohort flow plots, and streamline plots. These tensors may be rank 5 or larger, but the first few dimensions/indices are almost always related to whatever demographics we hope to study with that particular velocity grid

(e.g. the divide between those that receive free/reduced-price lunches and those that do not). The last four indices are always the same. In order, the indices correspond to the calendar year, the grade of the student, the "moments" of the student, and the score bins. In this context, "moments" refer to $N_{t,g,k}$, $v_{t,g,k}$, and $D_{t,g,k}$.[2]

Figure 3.2 shows an example of what a velocity grid might look like in a Python terminal. This example splits the TAKS data set along economic status, where a student is considered either economically well-off or receiving some kind of financial aid. Other file types may split the data along ethnicity, geographic region, campus number, or any combination of the above. In this example, we have chosen to look at students who are not receiving any financial aid who are transitioning from 6th grade in 2006 to 7th grade in 2007. Note that the second row corresponds to *total* velocities. Of the 39,309 students who scored between 90% and 100% on their 6th grade test, their grade went down by an average of $1677.6/39309 = 4.27\%$ from 6th grade to 7th grade. Also note that students that score in the 60%-70% score bin or higher tend to see a decreasing score change on average in the next year, while students below that tend to see an increase in score on average.

Retention grids are shaped identically to velocity grids, and in fact the only difference is that they do not measure "moments" or $v_{t,g,k}$. Instead, the penultimate index corresponds to two values: the number of students in a given

[2]The term "moments" arose from the fact that all three terms are related to $(s^{\alpha}_{t+1} - s^{\alpha}_{t})^{m}$, $m \in [0, 1, 2]$.

```
>>> velocity_grid = pickle.load(
                            open('VGDisVelGrid.pkl'))

>>> numpy.shape(velocity_grid)
(2, 9, 9, 3, 10)
# Dimensions are economic status (a binary variable),
# years, grades, moments, and score bins

>>> velocity_grid[0][3][3]
[8,   23,  135,  628,  1924, 3513, 8855,   12732,
27721,   39309]
[2.9, 6.2, 15.2, 52.8, 95.1, 89.1, -189.6, -550.3,
-1281.1, -1677.6]
[1.2, 2.9, 4.4,  16.7, 40.7, 61.9, 145.9,  202.2,
337.7,   257.9]
# Matrix corresponds to economically well-off students
# transitioning from 6th grade in 2006
# to 7th grade in 2007
```

Figure 3.2: Pseudo-code showing an example of velocity grid structure

year/grade/score bin, and the number of those who were in the same grade for the following year. *Exit grids* are also similarly shaped to both velocity grids and retention grids. In this case, the penultimate index corresponds to the number of students who disappeared from the TAKS data set from one year to the next, or the number of students who appeared in the set from seemingly nowhere. Retention grids are useful for seeing if students are being held back in great numbers to avoid high-stakes exams, and exit grids are useful for seeing if large numbers of students drop out before reaching those same exams.

Figure 3.3 shows pseudo-code that represents how velocity grids, retention grids, and exit grids are created. This particular example highlights a common case where students are divided by their economic status. Those that receive free lunches may be grouped together with those that receive reduced-price lunches as "economically disadvantaged", or they may be separated. Instead of economic status, we may choose to divide the students by their ethnicity, or by the region of Texas where they attend school, or by any combination of the above. All that changes are the first few dimensions of the grids, leaving the last four (years, grades, moments, score bins) essentially unchanged.

Trajectory files are CSV files where we group students according to the following: demographic information, initial grade, initial year, initial score bin, and future grade. For example, we may be concerned with all the low-income students who scored above 90% in 3rd grade in 2004. We would then have a

```
velocity_grid, retention_grid, exit_grid = numpy.zeros((
    number_of_economic_levels,
    number_of_years,
    number_of_grades,
    number_of_moments,
    number_of_score_bins,
    ))

for row in condensed_student_history_file:
    economic_status = row.extract_economic_info()
    all_grades = row.extract_grades()
    all_scores = row.extract_scores()

    for grade in all_grades:
        diff_score = score_next_year - score_this_year
        velocity_grid += diff_score**moment

        if grade_next_year == grade_this_year:
            retention_grid += 1
        if score_this_year is True \
        and score_next_year is False:
            exit_grid.student_disappearance()
        if score_this_year is False \
        and score_next_year is True:
            exit_grid.student_appearance()
```

Figure 3.3: Pseudo-code illustrating how velocity grids, retention grids, and exit grids are created

separate row for each of these students for 4th grade, 5th grade, and so on. Each of these rows shows the "moments" for that particular grouping, and as mentioned in the discussion on trajectory plots in Section 2.4.4, the number of students in each grade and their score changes are exact, not interpolated or assumed.

Figure 3.4 is an example of the code used to create trajectory files. The scores in "year1" and "year2" are translated into moments in a later step. Note that when looking at possible combinations of years, the first year does not need to come before the second. If we restrict ourselves to pairings where "year1 $<$ year2", that results in a *forward trajectory* and answers the question "Given an initial year, grade, and score, where did those students end up?" Reversing the inequality gives a *backward trajectory* and answers the question "Given a final year, grade, and score, where did those students come from?"

Markov files are useful for testing the effects of using multiple years of student test scores to predict future behavior. Their name comes from an older data type used by Marder and Bansal where students were counted based on their year, grade, their test score in that year/grade, and their test score in the subsequent year. If we fix both year and grade, the result is a matrix that looks exactly like a Markov chain matrix. It was this data type that led to the inspiration that characteristic eigenvectors might be able to be extracted from the data (see Section 5.2.4). We realized that we could extend this data type so that a given cell in the tensor would show us not just how students scored in two consecutive years, but how students scored in an arbitrary num-

```
for row in condensed_student_history_file:
    economic_status = row.extract_economic_info()
    all_grades = row.extract_grades()
    all_scores = row.extract_scores()
    year_combinations = [(2003, 2003),
                         (2003, 2004),
                         ...
                         (2004, 2003),
                         (2004, 2004),
                         (2004, 2005),
                         ...
                         (2010, 2011)]
    for year1, year2 in year_combinations:
        trajectories.record(
            economic_status,
            year1,
            grade_in_year1,
            score_in_year1,
            grade_in_year2,
            score_in_year2,
            )
```

Figure 3.4: Pseudo-code illustrating how trajectories are created

ber of consecutive years. Markov files were initially multidimensional tensors like velocity grids, but I discovered rather quickly that rank-8 and rank-9 tensors are not exactly memory efficient! Considering the number of zero values present in a large Markov file, we now use CSV files that only highlight the nonzero values. Figure 3.5 shows an example of how a fourth-order Markov file is created; note that the output step has a check so that the only rows written to the CSV file are those that will survive the FERPA masking. For each student, his or her scores from 2003 to 2007 consist of one data point, 2004-2008 is another, 2005-2009 is another, and so on. This is a fourth-order Markov file because each data point spans five years but four score transitions.

3.4 FERPA Constraints

The Family Educational Rights and Privacy Act of 1974, or FERPA, is a federal law designed to protect students at all levels of education. Specifically, it gives students (or parents, if the students are younger than 18) the right to control their education records and keep them private. FERPA is the law that protects students from, for example, having a teacher post their grades on a classroom wall. For the TAKS data set, great pains and measures have been taken to ensure that the data meets all the privacy requirements of FERPA. Unfortunately, those same regulations proved to be very frustrating to anyone who is accustomed to having free reign over his or her research material.

FERPA officially states that educational researchers are only allowed

```
order = 4
for row in condensed_student_history_file:
    economic_status = row.extract_economic_info()
    all_grades = row.extract_grades()
    all_scores = row.extract_scores()

    for all_possible_combinations:
        score_tuple = (
            economic_status,
            initial_year,
            initial_grade,
            initial_score,
            next_four_subsequent_scores,
            )
        output_dictionary[score_tuple] += 1

for score_tuple in output_dictionary:
    if output_dictionary[score_tuple] >= 5:
        write_to_csv(output_dictionary[score_tuple])
```

Figure 3.5: Pseudo-code illustrating how Markov files are created

access to educational data as long as "information will not permit the identification of any person by the organization receiving such information." The fact that all students are referred to by an anonymous identifying number does not satisfy this requirement; a very small school might only have a single student that identifies as American Indian, and that student would be easily identifiable in the data set. For Texas and the TEA, this FERPA statute has been interpreted to mean that researchers should not be able to isolate fewer than five students at a time. If any cell of the data has a student count of less than five, or if any cell has information that has been derived from fewer than five students, that cell needs to be "masked" in such a way that the researcher is unable to determine the values of those cells. Masking can take various forms, but in many cases it is sufficient to simply set all offending cells to zero.

To verify that all data that comes out of the ERC is FERPA-compliant, people have been set in charge of reviewing it and looking for any cells that are less than five (or ones that may be deduced from surrounding cells). When I began working on this project, I was informed that these reviews were done by hand on printouts of the data. This would prove difficult for our data, as it exists mostly as tensors of at least rank-4 instead of SAS or STATA spreadsheets. Furthermore, we use the Python module "pickle" in order to save the files, which makes them unreadable to those who do not have access to a Python interpreter or the pickle module. Consequently, I have written programs that will take our tensors and slice off individual matrices into a CSV file for TSP review. The TSP has agreed to release our pickled files to us on

our assurance that the information contained in the pickled files is identical to that of the CSV files. Unfortunately, given that our tensors can easily exceed millions of cells, the review process is often quite lengthy. Weeks may pass between submission of our data for FERPA review and the release of that data to us.

FERPA also constrains us from excessive granularity. When creating a high-rank Markov file, I may know nothing about a student except the year he or she was in 3rd grade, and his/her score percentile for every year after that. If that student has a unique set of those values, then that record is struck from our data. This masking is irrelevant when looking at statewide populations; 27 million student records are spread across a few thousand cells, and masking eliminates less than 1% of the data set. Masking is a much bigger problem when examining high-rank Markov files (each rank increases the granularity by a factor of 10) or when trying to isolate individual campuses. Only the largest schools in the state have enough students so that the masked data still contains valuable information. Marder and Bansal initially gained access to a publicly available version of the TAKS data set, but that set had only 14 million records, and over 30% of those had been masked and their values replaced with asterisks.

Finally, the TEA's worries about data security have led to some peculiar technological constraints. Our contract to use the TAKS data set was granted through the University of Texas at Dallas, and the data is stored on servers at Dallas. The TSP had an office in Austin through which one could remotely

access the data in Dallas (although those computers were isolated from the rest of the Internet). However, this office was eventually shut down, and the data set has been completely sequestered. The only way to access the data now is by working through one of the computers that is physically located at UT-Dallas. Thumb drives are not permitted, although the staff at the ERC are able to put pre-coded scripts on their servers for researchers.[3] Some coding may be anticipated and finished ahead of visits to the ERC, but inevitably any trip includes several hours of fixing bugs and rewriting code.

As of December 2013, the ERC at UT-Dallas does not have a systems administrator in their employ. Since the entire system is sequestered from the Internet and only accessible through a few systems, this is an acceptable situation since there is no need for patches or security updates. However, this means that much of the software they use, while operable, may be outdated. Code that was written with Python 2.7 may break when using Python 2.4.3 on the ERC's systems.

[3]It must be emphasized again that the staff at the ERCs in both Dallas and Austin were extremely helpful in working with us to overcome these hurdles and to accommodate our unique methods/data structures. Conducting our research would have been a much more difficult task without their help.

Chapter 4

Results

4.1 Changing the Flow

Marder and Bansal showed that flow plots can change, and often quite dramatically, as students are separated based on their economic status [31]. The initial goal of my research was to find further instances where separating students by various demographic variables (e.g. gender, ethnicity, campus) would lead to statistically significant differences in their corresponding flow plots. Unfortunately, many of those variables do not split the population evenly enough to make comparative statistics useful. It was only coincidence that dividing the state of Texas along economic lines resulted in a nearly 50%-50% split, and even isolating ethnicity makes it difficult to uncover useful information about the American Indian and Asian populations.

Another early goal was to extend the methods developed in Marder and Bansal. While snapshot flow plots are useful, a common mistake in interpreting them is to forget that each column of arrows corresponds to an entirely different population of students (see Section 2.4.1). In contrast, cohort flow plots are more useful for longitudinal studies but take much longer to generate. Once

we had been given access to seven years of unfiltered test data, we could create substantive cohort flow plots that compared and contrasted the performance of different cohorts over more grades than before. We did observe changes in the score distributions of different cohorts, but the nature of the changes was unusual.

Each column of a flow plot represents a distribution of student scores, and the area of the arrows show how that distribution is aligned. Starting with the graduating class of 2012, there is a sudden change in the distribution of scores at 5th and 8th grade; this is shown by comparing the cohort flow plots in Figure 4.1. Whereas prior cohorts display a relatively uniform score distribution at all grades (with the average of the distribution being lower for economically disadvantaged kids), the test scores of the class of 2012 are remarkably top-heavy at 5th and 8th grade. The effect was not isolated to a single cohort either; the class of 2013 showed similar effects. We will examine these changes in more detail in Section 4.3.

It became clear to us that something in the system had changed, but the cause of the change was initially a mystery to us. As mentioned in Section 1.2.1, 5th and 8th grade are high-stakes testing years, but we did not see a similar top-heavy distribution for 11th graders. We considered possible sources of students that might have suddenly skewed the score distributions; for example, in 2008 the State-Developed Alternative Assessment II (SDAA II) was terminated. Prior to 2008, all students taking the SDAA II did not have their tests scored, and their results were not considered in determining compliance

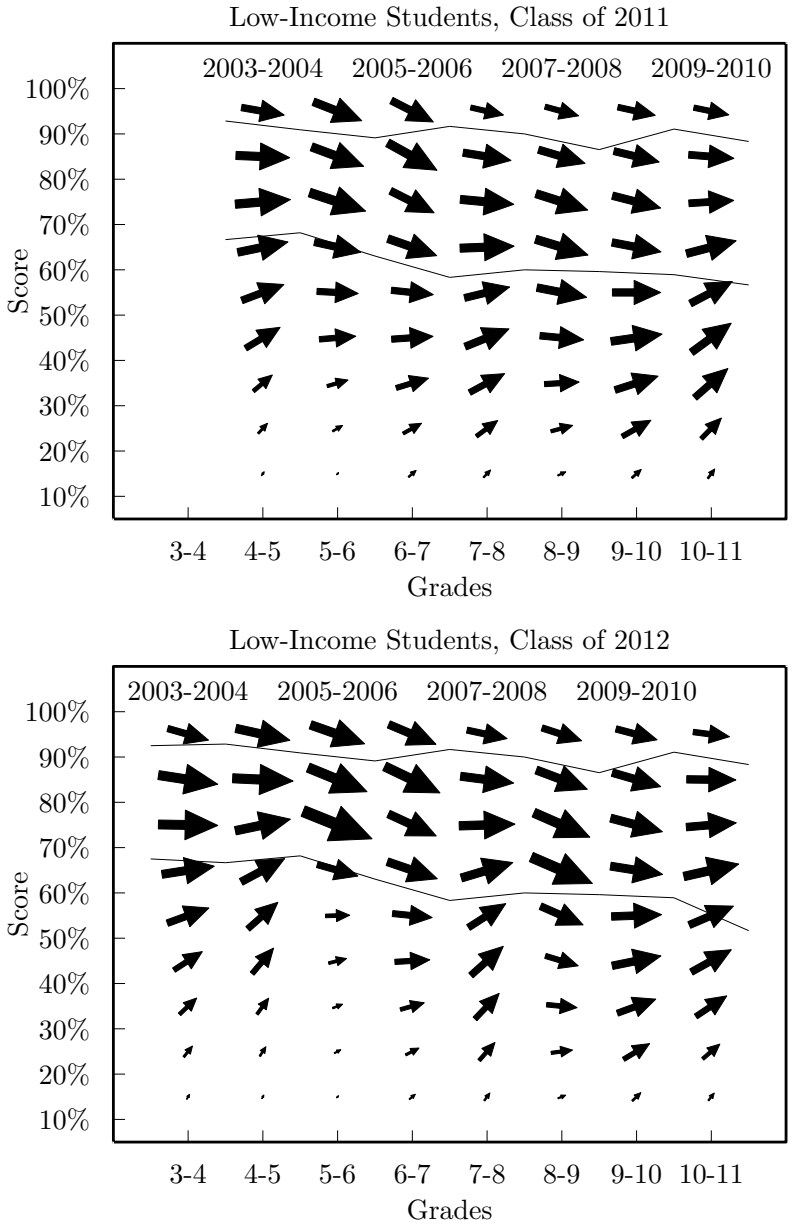

Figure 4.1: An example of different cohort flow plots comparing the classes of 2011 and 2012

with NCLB. This had the effect of introducing approximately 60,000 students into the testing population, but that number is miniscule compared to the millions of students taking the test each year. Furthermore, the SDAA II was still an option for the class of 2012 when they were 5th graders in 2005.

4.2 Student Success Initiative

In 1999 the 76th Texas Legislature approved funding for the Student Success Initiative (SSI), a policy which targeted those students who struggled with the reading and math tests in the high-stakes years of 3rd, 5th, and 8th grade. It was the initial SSI that provided the legislative framework of the high-stakes years, namely that students would be required to pass the 3rd grade TAKS reading test to advance to 4th grade, and students would need to pass the reading and math tests in 5th and 8th grade to advance to 6th and 9th grade respectively [49].[1]

In order to target underperforming students in high-stakes grades, the SSI directed the majority of its funding to the Accelerated Reading/Math Instruction grant programs (ARI/AMI). These funds would allow districts to provide additional tutoring and instruction to students who failed the TAKS subject tests in high-stakes years. In addition, students would be allowed to retake the TAKS test up to two more times. The students would only be

[1] The 81st Texas Legislature later removed the requirement that 3rd graders would need to pass the reading TAKS test before advancing to 4th grade.

retained a grade if they failed the TAKS test on their third attempt. Interestingly, no provisions were made for those students who failed the science test in 5th grade or the science and social studies in 8th grade, as these subject tests were not required for grade advancement.

The funding that districts received for ARI/AMI was directly tied into the performance of students on the high-stakes examinations. For the 2006-07 school year, school districts were awarded $1,548 for each 3rd and 5th grade student who failed the 2006 reading and math TAKS tests, respectively [45]. These funds were to be used to help all underperforming students up to a certain grade level (see Table 4.1), but the number of students being served kept increasing while the funding leveled off in the 2004-05 school year. The initial class of kindergarteners served by the ARI program totaled 75,340 students, while almost 1.2 million students were served by ARI and/or AMI in the 2006-07 school year [45]. Consequently, the average funding that each student received went from $320/student in the 2000-01 school year to $120/student in the 2006-07 school year.

In 2007, the 80th Texas Legislature decided to move the funding away from the student-focused ARI/AMI towards statewide teacher professional development programs. In 2009 the 81st Texas Legislature transitioned the ARI/AMI program into the Student Success Initiative Grant program (SSIG). The SSIG continues to fund additional tutoring for students, but its budget was slashed by approximately a factor of three (see Table 4.1). SSIG does grant districts more flexibility on how funding may be spent; instead of focusing

Year	Funding (M$)	Grades	Year	Funding (M$)	Grades
99-00	65.2[a]	K	05-06	149.5	K-6
00-01	57.5[a]	K-1	06-07	144.2	K-7
01-02	106.4[a]	K-2	07-08	124.9[e]	K-8
02-03	75.1[a,b]	K-3	08-09	123.3	K-9
03-04	80.9[c]	K-4	09-10	44.2[f]	K-12
04-05	144.1[d]	K-5	10-11	44.4[f]	K-12

Table 4.1: Funding history of ARI/AMI. (a) Accelerated Reading Initiative (ARI) funding only (b) First year grade 3 had to pass (c) Accelerated Mathematics Initiative (AMI) funding begins (d) First year grade 5 had to pass (e) First year grade 8 had to pass (f) ARI/AMI defunded; Student Success Initiative Grant only [49]

solely on reading and math in pre-high school grades, funds may now be spent on any subject in any grade.

A unique feature of the ARI/AMI program was the manner in which it was rolled out, and how that schedule affected the graduating class of 2012. The first year of ARI/AMI was the 1999-2000 school year, and the funding was only allowed to serve kindergarteners who were underperforming in reading. The 2000-2001 school year allowed for accelerated instruction in reading for kindergarteners and 1st graders, the 2001-2002 school year targeted K-2 students, and so on. Math was not addressed until the 2003-2004 school year, when both ARI and AMI became active for students between kindergarten and 4th grade. Every successive year allowed for a new grade to be served by additional tutoring and instruction up until the defunding of the program in 2009.

This rolling implementation meant that in every year of the ARI/AMI

program, the new grade being served was part of the graduating class of 2012. Furthermore, it was the class of 2012 that initially encountered each of the high-stakes tests. 3rd graders did not have to pass the TAKS reading test to graduate until 2003, 5th graders did not have to pass TAKS reading/math until 2005, and 8th graders did not have to pass TAKS reading/math until 2008. There is a clear dichotomy between the cohort of 2012 and all the cohorts before it; the class of 2012 was the first to face the hurdles of high-stakes testing in every grade, but they were also the first to receive funds for additional tutoring and instruction in every grade. It was this dichotomy that stood out to us as we examined the TAKS data.

4.3 The Effects of ARI/AMI

4.3.1 Cohort Plot Revelations

Figures 4.2 and 4.3 show the snapshot flow plots for the 2003-04, 2005-06, and 2008-09 transitions for both economically well-off and disadvantaged students; these are the years that the effects of high-stakes testing would first be noticed. Note how the arrows above the passing cutoff line in the 5th to 6th grade transition are much larger beginning in 2005-06. Also, the effect is stronger for economically disadvantaged students, despite the fact that a larger portion of their population was failing in the 2003-04 transition. Similarly, the arrows for the 8th to 9th grade transition follow a similar trend beginning in 2008-09. This was our first evidence that something unusual was occurring

in our flow system; a perturbation had occurred, and we did not know what had caused it. We considered several alternative explanations for this shift, including the possibility that low-scoring students were being pushed out of the data set to take the State-Developed Alternative Assessment II (SDAA II). The SDAA II is a version of the TAKS that special education students are allowed to take; however, our examination of the data showed that students were not dropping from the data set at any greater rate than they had been prior to the perturbation taking place. Furthermore, it is not likely that the test suddenly became easier in the high-stakes testing years. As mentioned in Section 1.2, questions are assigned difficulty levels through a proprietary process, and they are selected to create a standardized test of approximately the same difficulty as in prior years.

Figures 4.4, 4.5, 4.6, and 4.7 show several cohort flow plots for the classes that would graduate between 2010 and 2013. Comparing the 2010-11 graduating classes to the 2012-13 graduating classes strongly indicates that the perturbation begins with the class of 2012 and continues for future cohorts. The timing of the perturbation and the grades affected by it led us to believe that what we were seeing was a direct consequence of the SSI and specifically the Accelerated Math Initiative. When given the option to retake the test with additional instruction and tutoring, the number of students who passed the TAKS mathematics exam in 5th and 8th grade was much higher than in cohorts prior to the class of 2012.

To further illustrate the differences between the class of 2011 and the

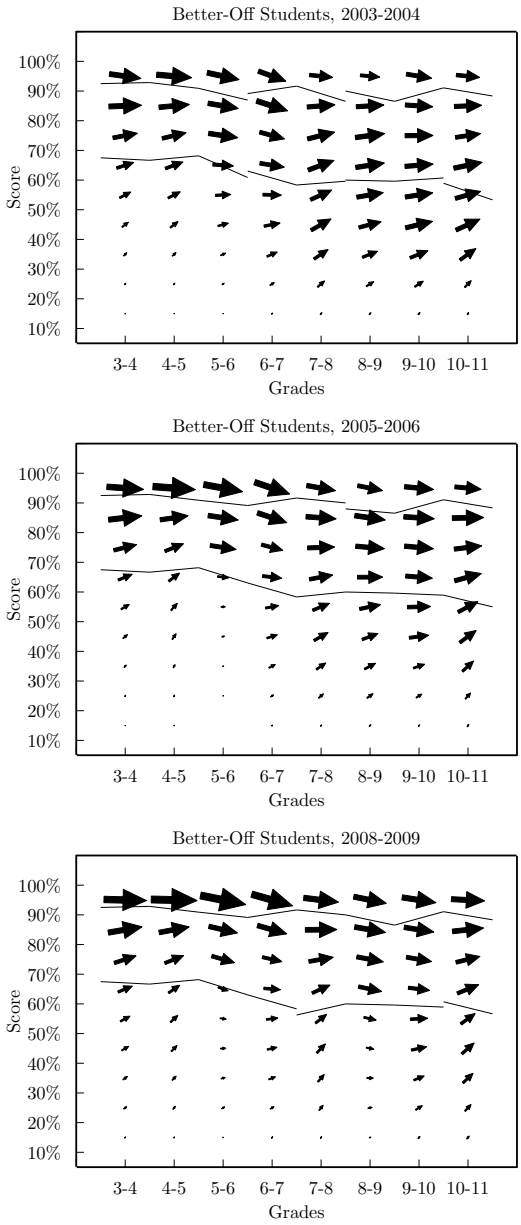

Figure 4.2: Snapshot flow plots for economically well-off students from 2003-04, 2005-06, and 2008-09

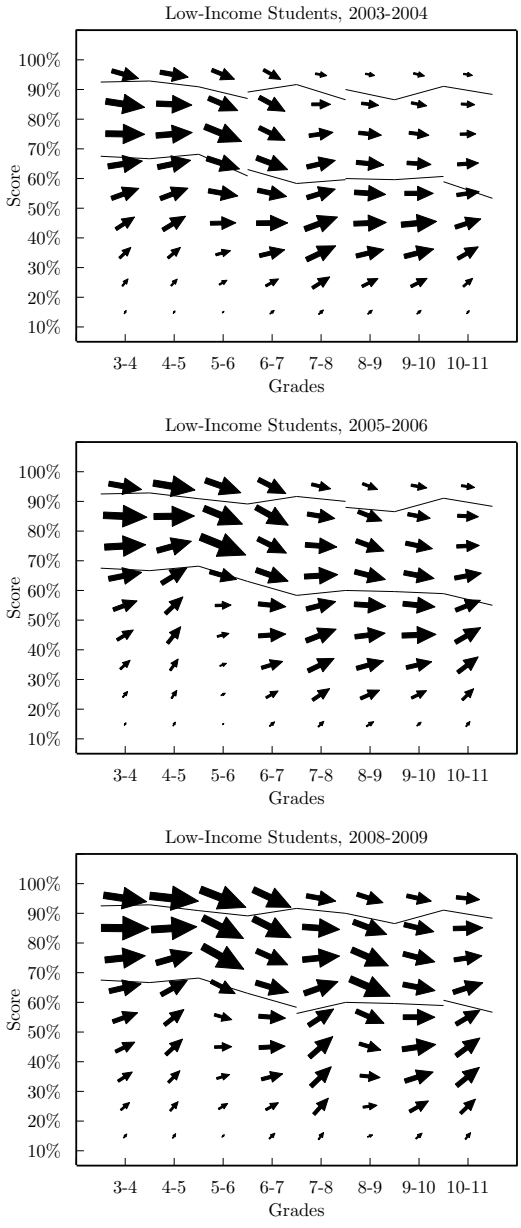

Figure 4.3: Snapshot flow plots for economically disadvantaged students from 2003-04, 2005-06, and 2008-09. Note how the 5th to 6th grade transition becomes significantly more concentrated above the passing cutoff line between the 2003-04 snapshot and the 2005-06 snapshot. A similar effect occurs for the 8th to 9th grade transition between 2005-06 and 2008-09.

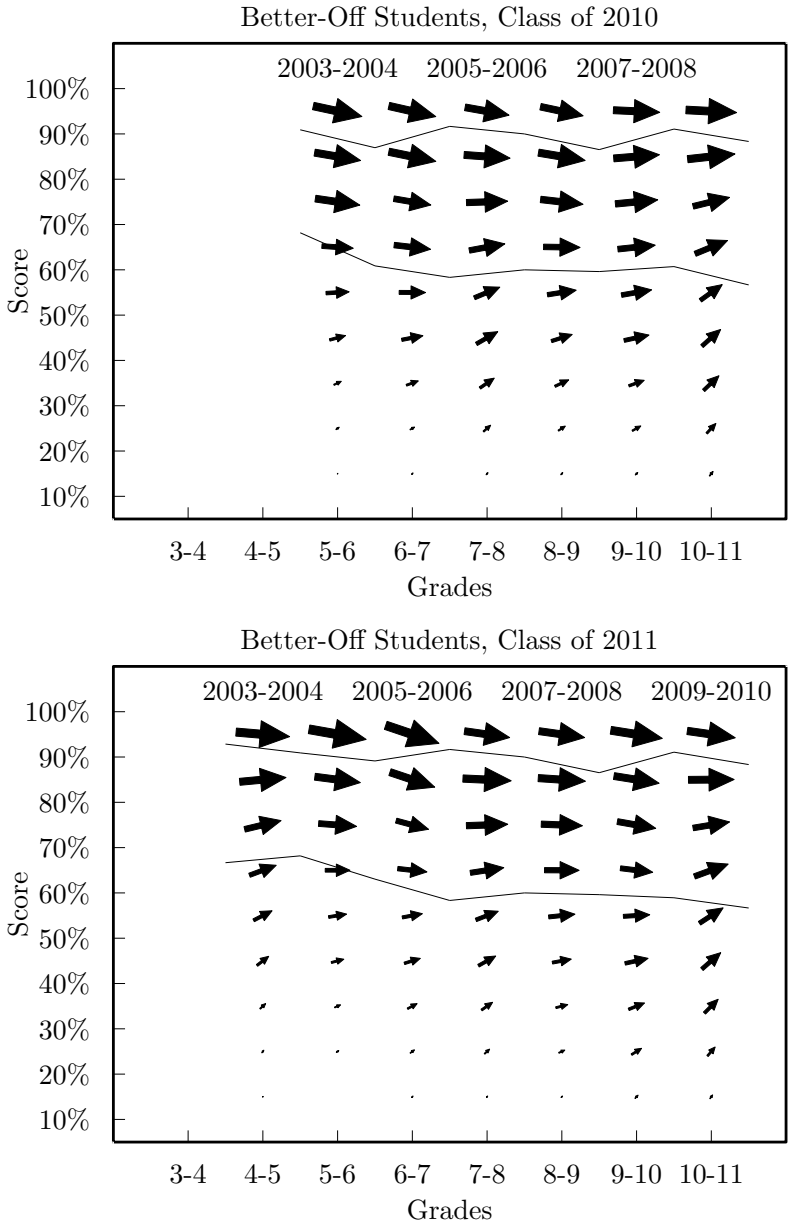

Figure 4.4: Cohort flow plots for the economically well-off students graduating in 2010 and 2011

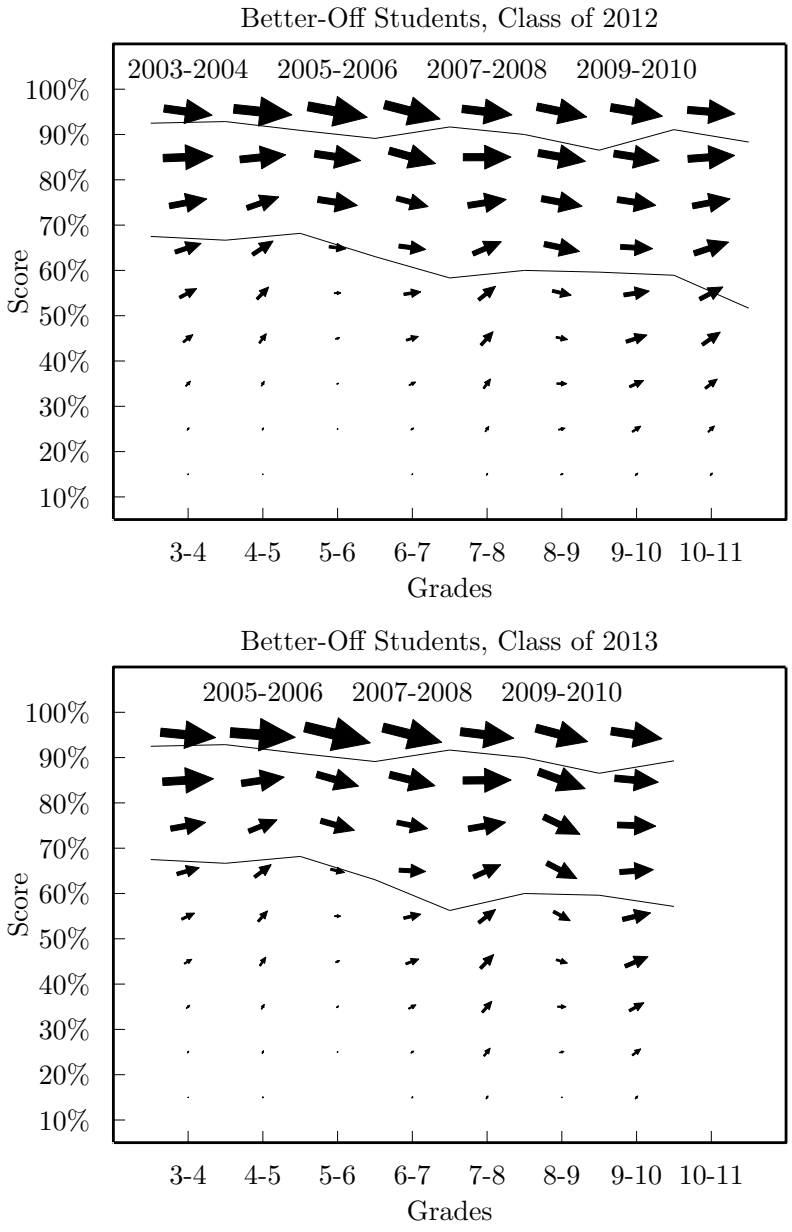

Figure 4.5: Cohort flow plots for the economically well-off students graduating in 2012 and 2013. Compare to Figure 4.4; the arrows below the cutoff line in the 4th-5th and 7th-8th transitions point much higher in this figure. Also, the arrows above the cutoff line for the 5th-6th and 8th-9th grade transitions are larger, indicating that more students passed the TAKS math test in those years.

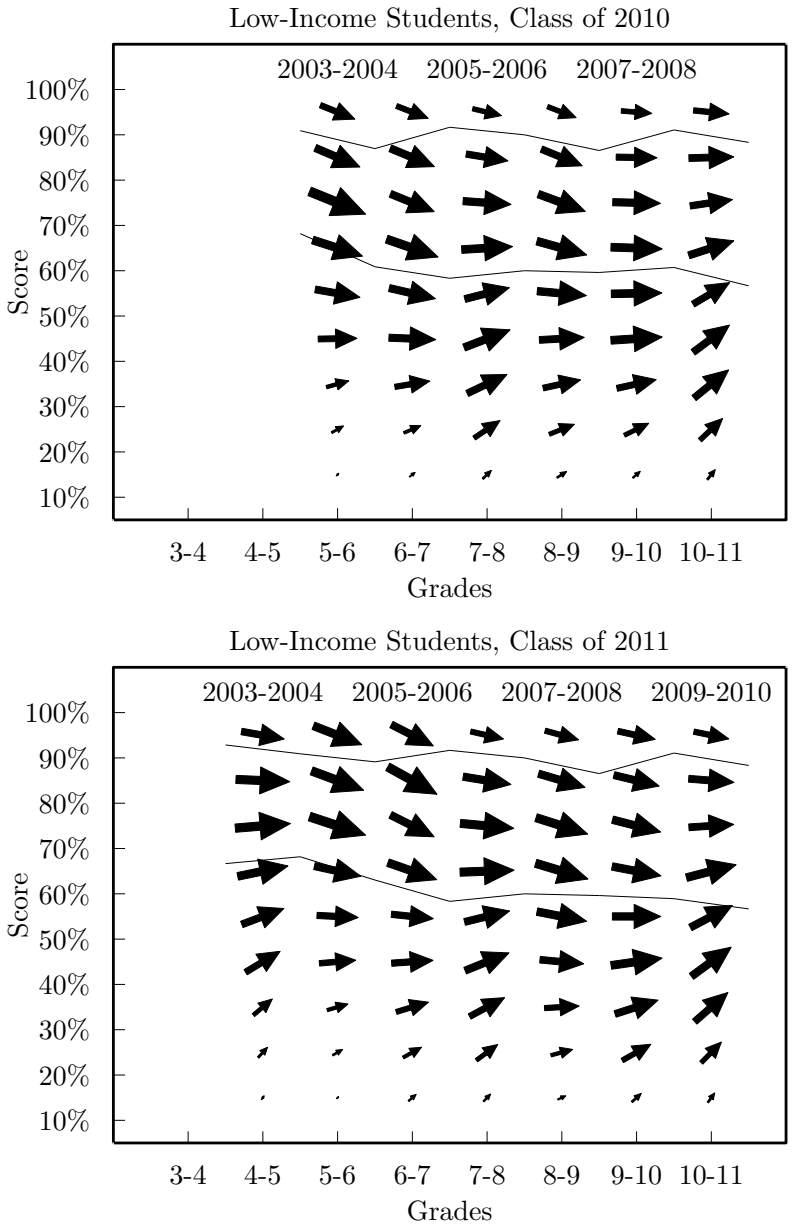

Figure 4.6: Cohort flow plots for the economically disadvantaged students graduating in 2010 and 2011

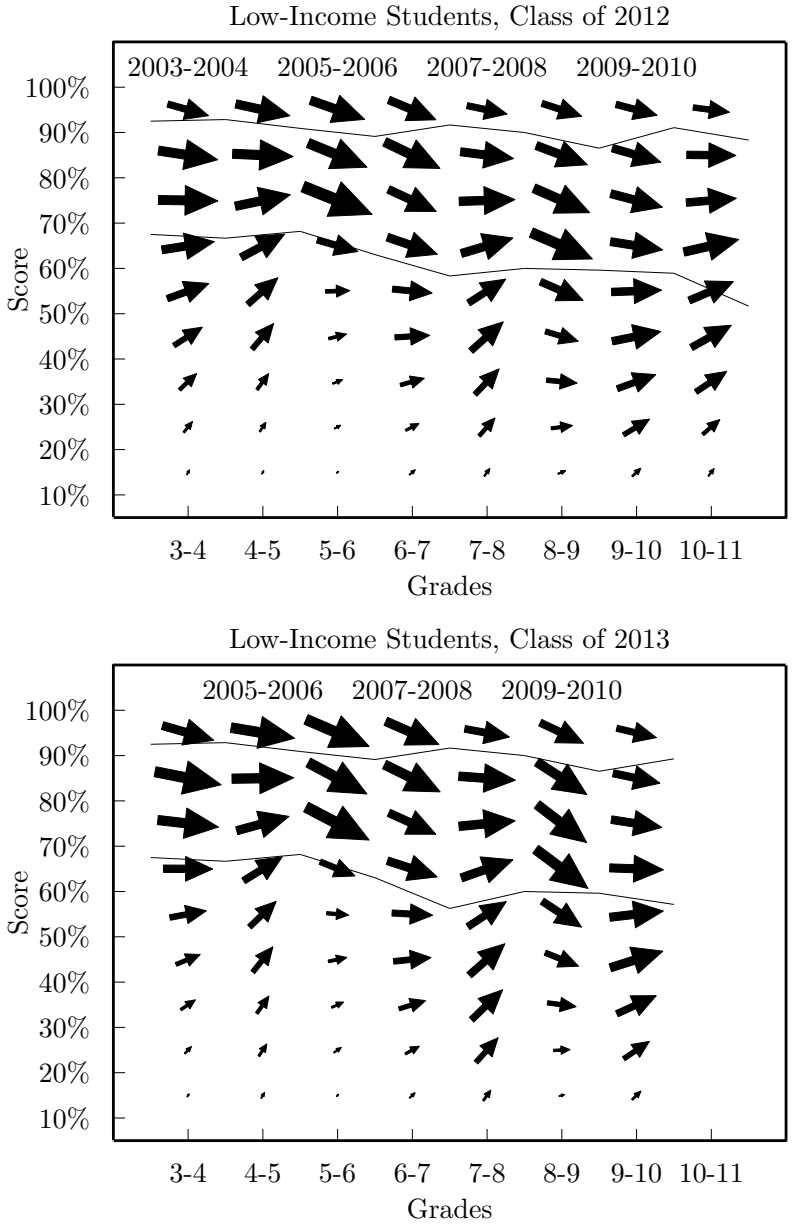

Figure 4.7: Cohort flow plots for the economically disadvantaged students graduating in 2012 and 2013. Compare to Figure 4.6; the arrows below the cutoff line in the 4th-5th and 7th-8th transitions point much higher in this figure. Also, the arrows above the cutoff line for the 5th-6th and 8th-9th grade transitions are larger, indicating that more students passed the TAKS math test in those years.

class of 2012, those cohorts are reprinted in Figure 4.8 but with coloring that indicates the relative gain or loss between the cohorts. A green arrow in the class of 2012 indicates a larger gain or smaller loss in score when compared to the class of 2011, and a red arrow indicates a smaller gain or a larger loss. (We do not reproduce the arrows of the 3rd to 4th grade transition for the class of 2012 here, since they have no analogue in the class of 2011.) We note in Fig. 4.8 the strong green coloring of arrows below the 80th percentile for the 4th to 5th grade and 7th to 8th grade transitions. This indicates that the class of 2012 saw much larger gains between those years than the class of 2011. However, we also note that the arrows for the 5th to 6th grade and 8th to 9th grade transitions are mostly red. The large gains experienced by the class of 2012 are soon followed by large losses. The question of whether the losses completely offset the gains will be addressed in Section 4.3.2.

In addition to separating students by their economic status, we also considered dividing them by their reported ethnicity to see if the plots significantly differ from one another. Figure 4.9 shows cohort plots for the classes of 2011 and 2012 for selected ethnicities.[2] We can see that, despite differences in population sizes, all three ethnicities show patterns similar to the economically disaggregated graphs. The class of 2011 shows arrows that slowly vary from one year to the next, while the class of 2012 show sharp increases for students below the passing line cutoff from 4th to 5th grade and from 7th to 8th grade.

[2]We do not display the plots for American Indian or Asian students because their populations are so small for the state of Texas that it is difficult to produce useful statistics or graphs.

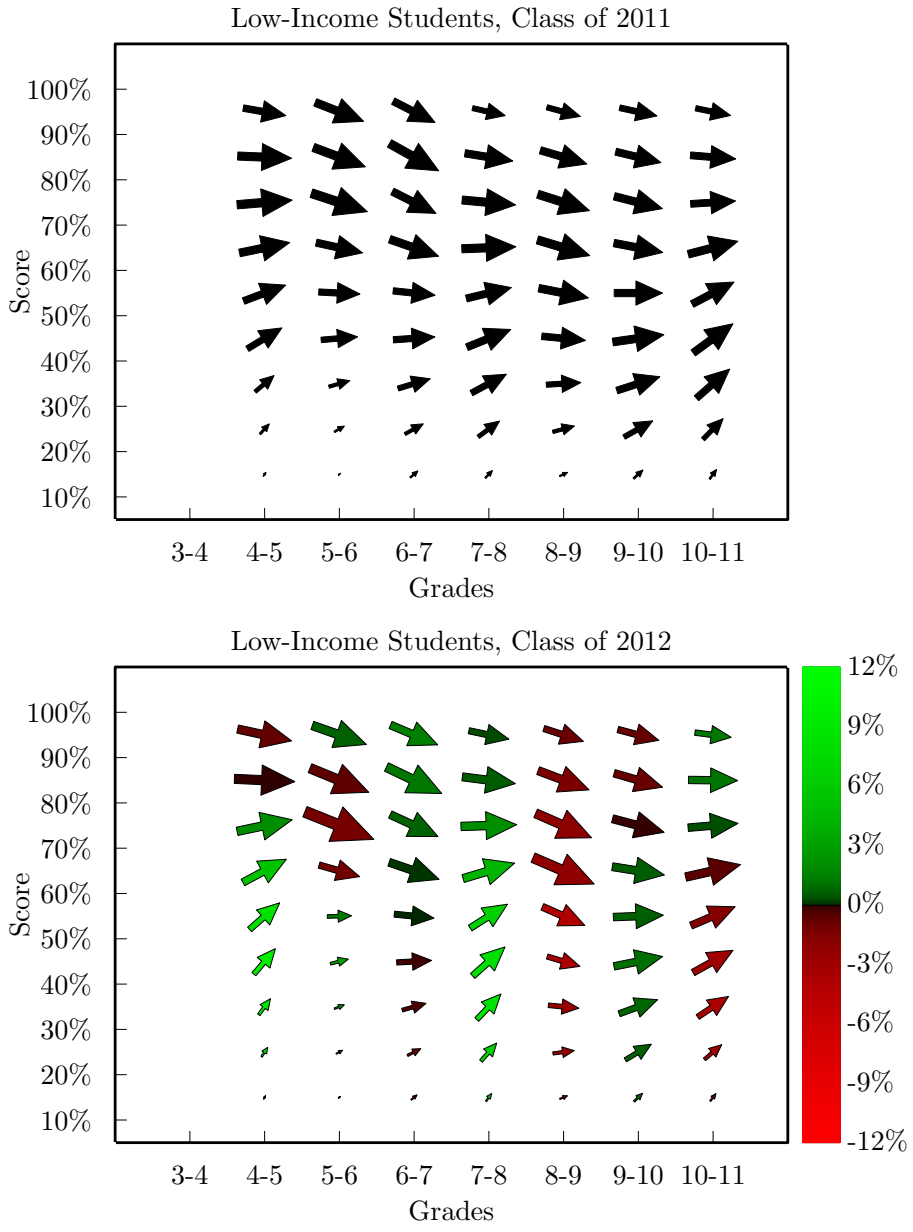

Figure 4.8: Colored cohort arrow plots directly comparing the class of 2011 and the class of 2012. Green indicates a larger gain or smaller loss in score when compared to the class of 2011, and red indicates a smaller gain or a larger loss. While gains in 5th and 8th grade are offset by losses in 6th and 9th grade, the net effect is positive; see the discussion of trajectories in Section 4.3.2.

We can even take this one step further and divide students by both their ethnicity and their economic status. Figures 4.10 and 4.11 compare several cohort plots, and the pattern remains the same. Note, however, that the arrow sizes are becoming very small as we disaggregate our data set further.

It has to be mentioned that this result was observed partially due to the method of data selection/refinement. We selected only one test score for a given year for each student, and we always selected the highest score that they attained (see Section 3.2). This standard was initially adopted by our research group to make it easier to create snapshot flow plots; if each student has a single data point in each year, the process of averaging score changes to create arrows is greatly simplified. However, this means that students who retake the tests will not have their failing test scores counted. It is possible that if we only looked at the first administration of each test, the flow plots of the class of 2012 might resemble those of the class of 2011 and before. Fortunately, we discovered that the effects of ARI/AMI are much more persistent than a simple data artifact.

4.3.2 Streamlines and Trajectories

Figures 4.12 and 4.13 compare the streamlines and trajectories, respectively, of the class of 2011 and the class of 2012.[3] In both cases, the plots peak sharply in 5th and 8th grade for the class of 2012. Figure 4.14 separates the

[3]Despite the fact that we have data for 3rd graders in the class of 2012, the plots all start in 4th grade so that we use identical initial conditions.

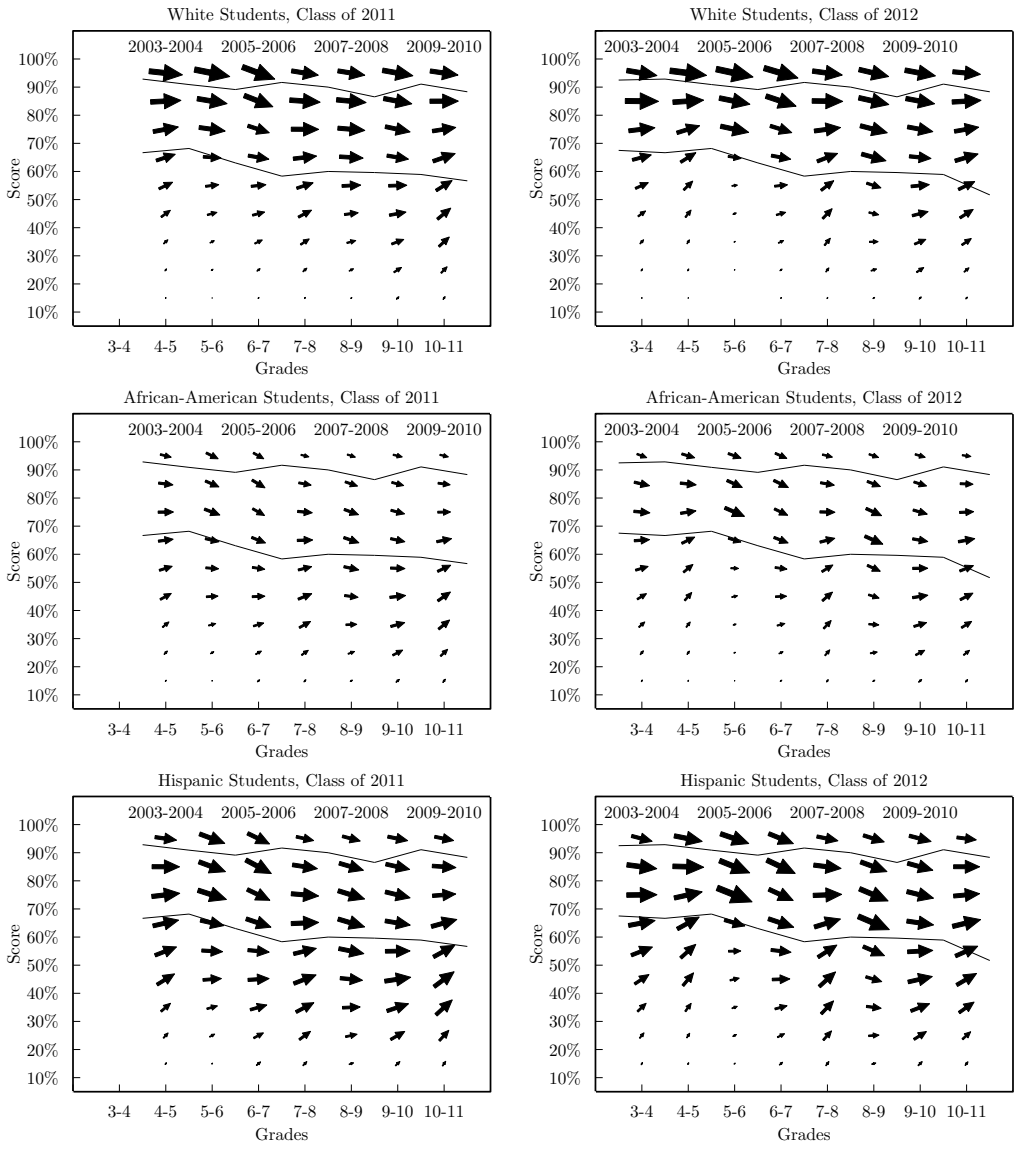

Figure 4.9: Side-by-side comparison of the classes of 2011 and 2012 for selected ethnicities

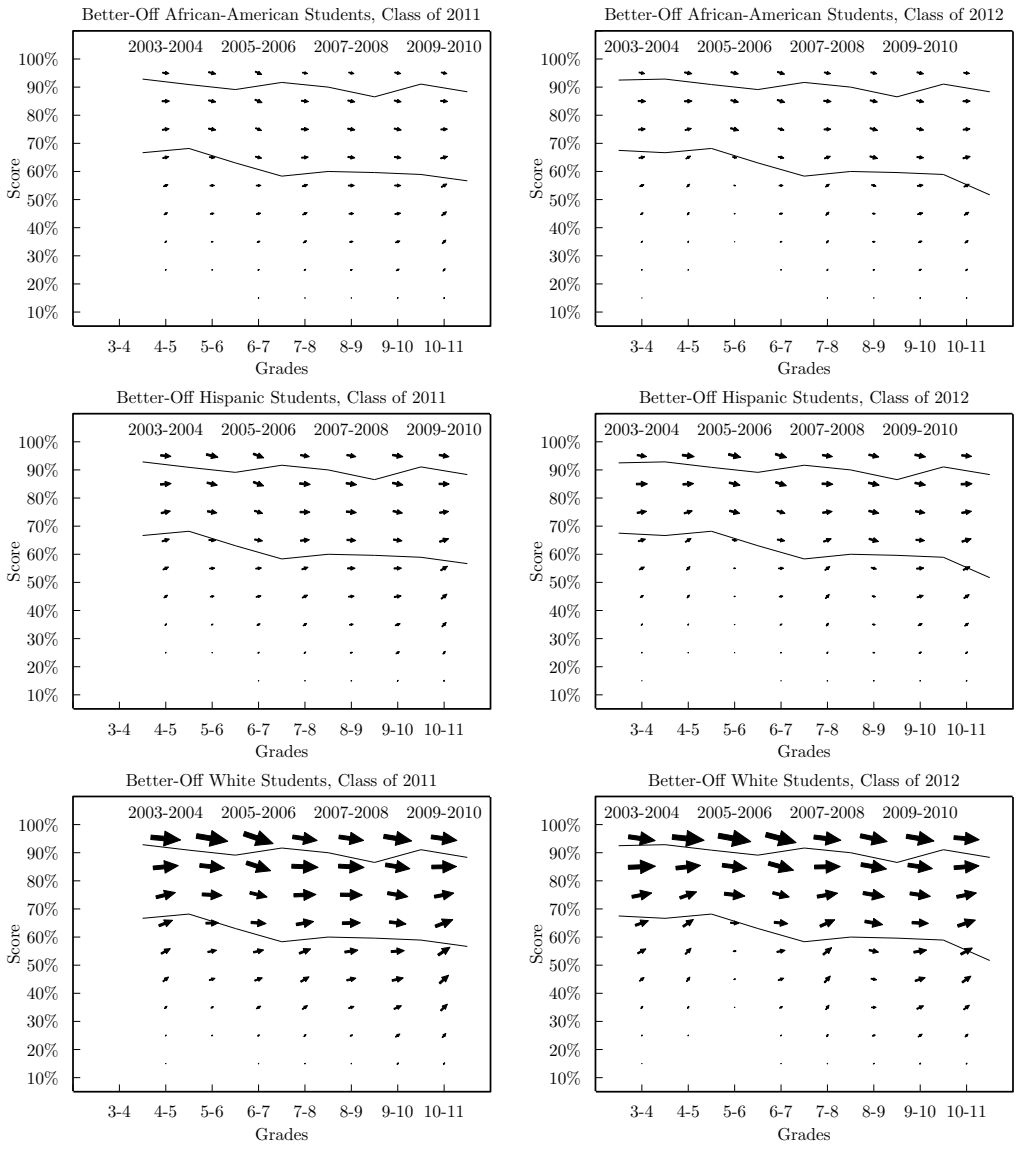

Figure 4.10: Comparing the classes of 2011 and 2012 for selected ethnicities, better-off students

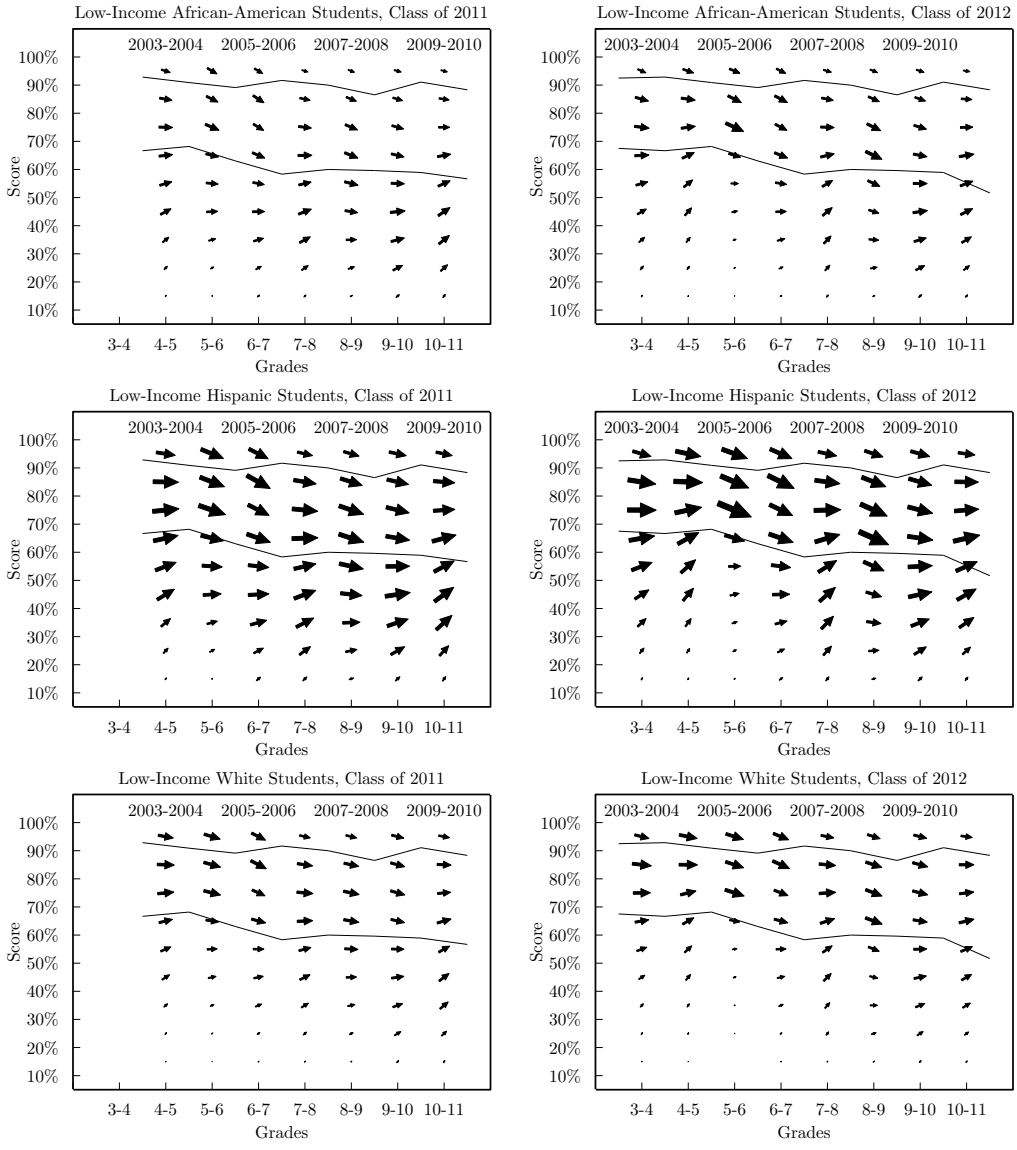

Figure 4.11: Comparing the classes of 2011 and 2012 for selected ethnicities, low-income students

students based on their economic status and directly compares the trajectories of the classes of 2011 and 2012 on the same graph; the peaks in 5th and 8th grade are even more noticeable here. This effect is most noticeable for those who scored in the 60%-70% score bin and below, which supports the claim that ARI/AMI was targeted at those who failed the initial high-stakes test. Note that the effect is prevalent for both low-income students and those that were not receiving any kind of financial aid. The scores of the low-income students do not quite rise to the same levels as the better-off students, but they only lag behind by a couple of percentage points in 5th grade. This is an especially remarkable result when considering that Figures 4.12 and 4.13 are referring to the entire statewide population of Texas; the thickest lines in these figures represent tens of thousands of students.

The surprising and striking revelation of Fig. 4.14 is what happened to AMI-affected students in grades that are not considered high-stakes. In 6th, 7th, 9th, and 10th grade, students who had been given additional tutoring and instruction through AMI outperformed their counterparts from the class of 2011, despite the fact that the tests are designed to have the same difficulty level and a similar cutoff line for passing scores. Those grades do not allow the option of retaking the test, so there is no bias from data selection. Furthermore, as shown in Table 4.1, the funding for ARI/AMI was cut by the time the class of 2012 reached 10th grade. When we look at the explicit traces of student scores in Figure 4.14, we see that the same trend of non-high-stakes improvement emerges regardless of economic status.

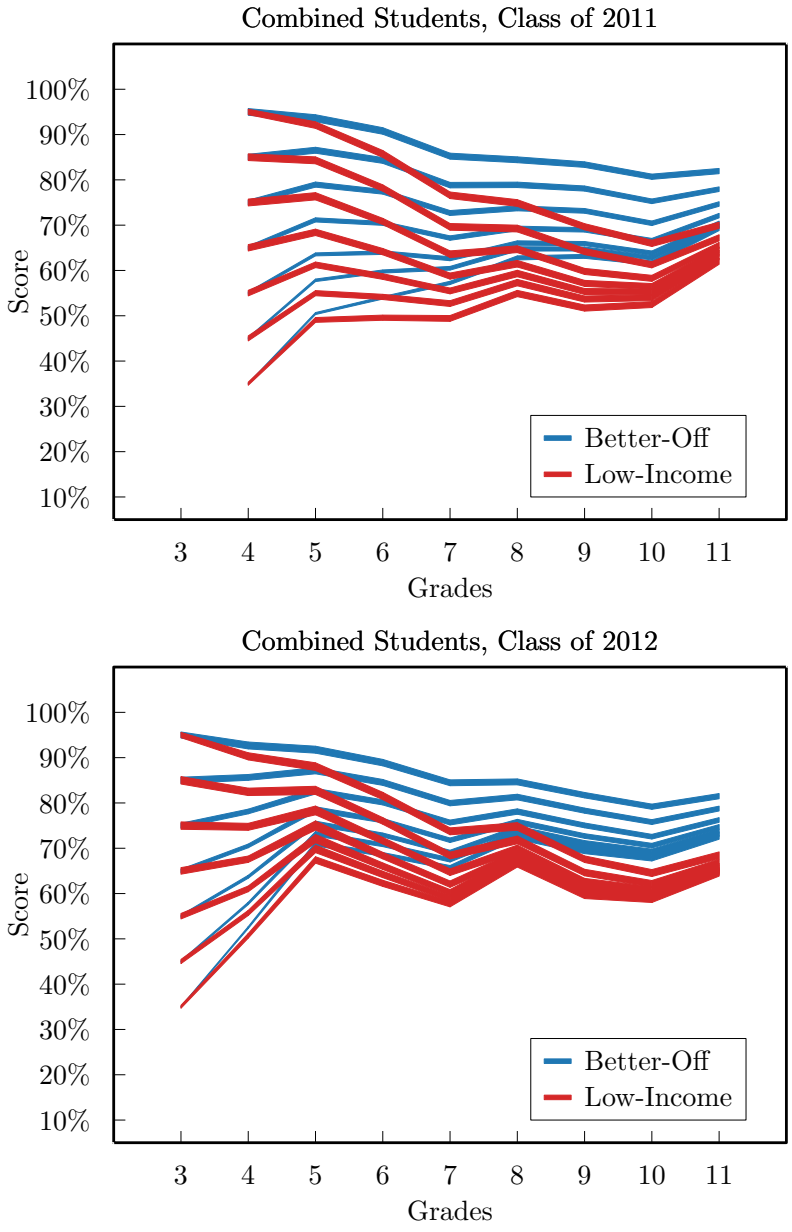

Figure 4.12: Streamline plots comparing the class of 2011 to the class of 2012. The average scores in 11th grade for all streamlines are higher for the class of 2012 than the class of 2011, and this is true regardless of economic status.

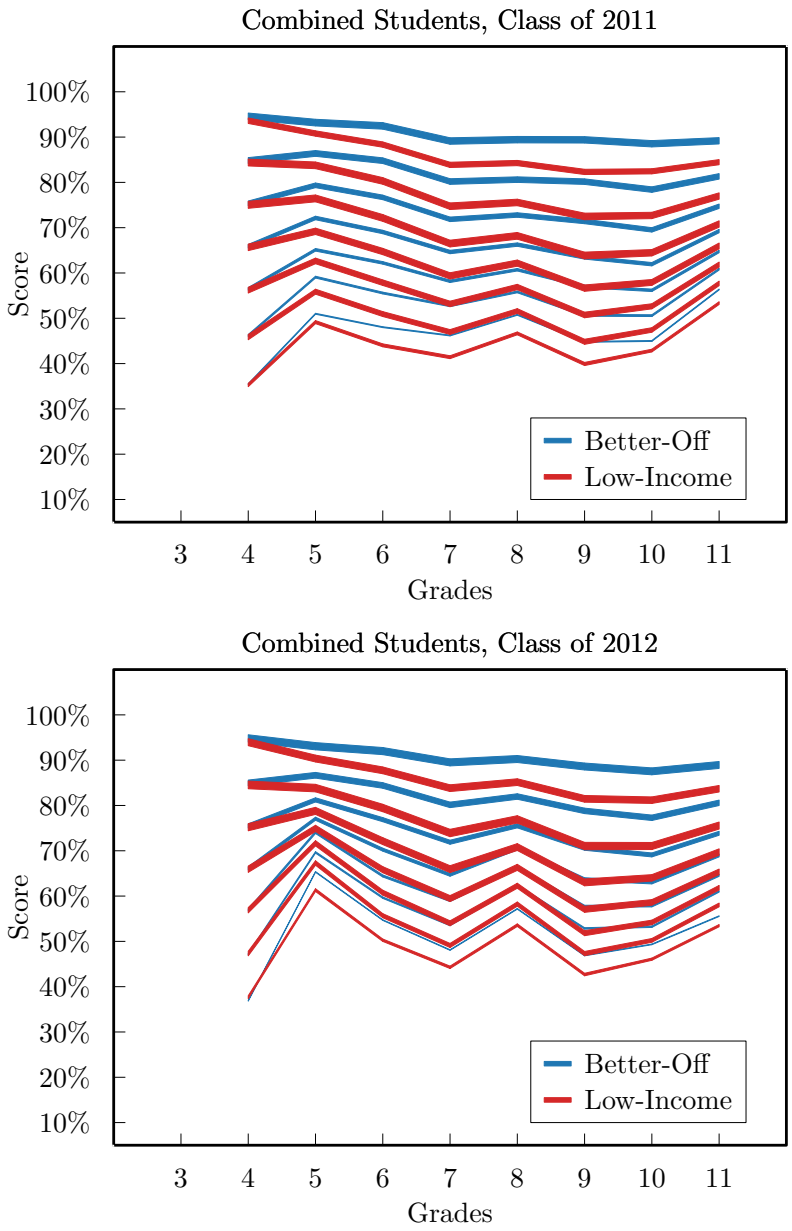

Figure 4.13: Trajectory plots comparing the class of 2011 to the class of 2012

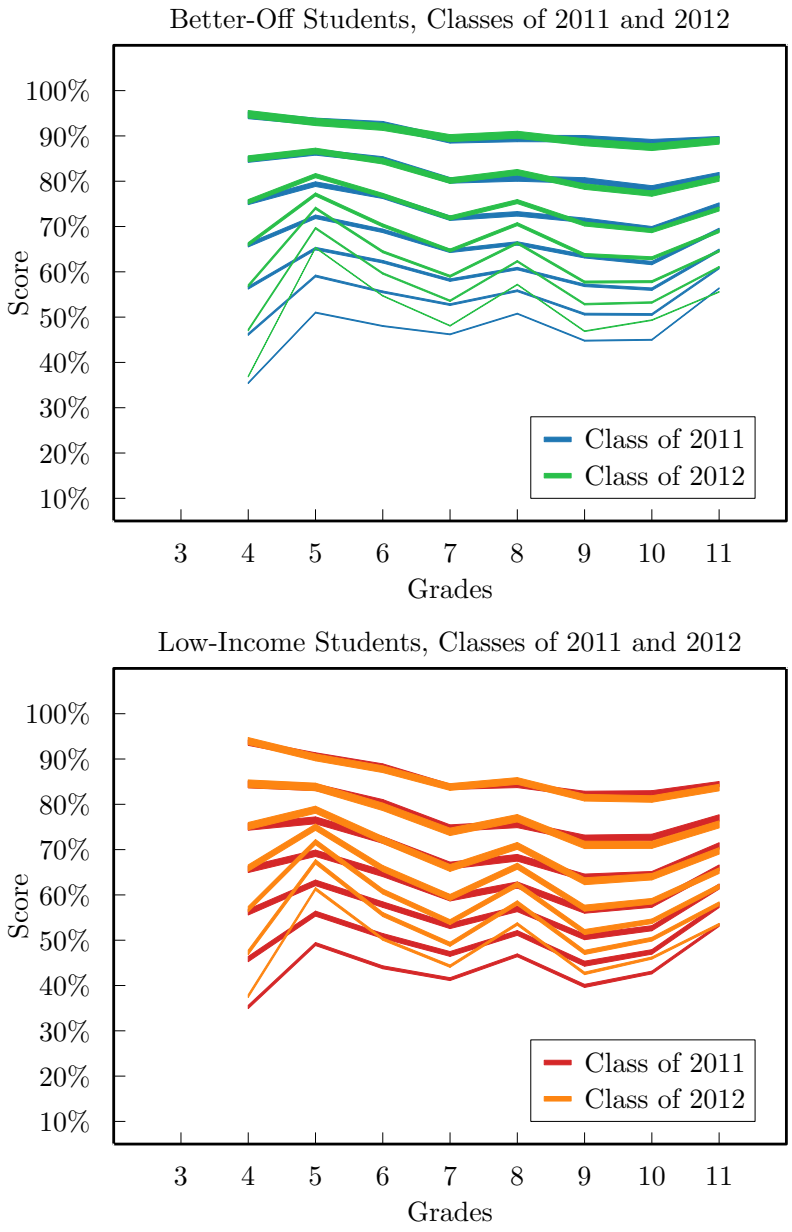

Figure 4.14: Trajectory plots comparing students of similar economic status from the classes of 2011 and 2012

The significance of these score gains should not be understated. Hanushek and Rivkin have produced research on how the quality of teachers may improve test scores, and to what degree. They estimate that a teacher from the 75th percentile of a quality distribution will produce learning gains of approximately 0.2 standard deviations over a teacher from the 25th percentile of the quality distribution [26]. These 0.2 standard deviations are with respect to some type of achievement distribution; the TAKS scores may serve this purpose here. For this standardized test, a gain of 0.2 standard deviations corresponds roughly to a 3%-4% increase in test scores. Figures 4.13 and 4.14 show that students in the class of 2012 who were initially low-performing made gains of this order in every year from 5th grade onward.

The high-performing trajectory lines show differences as well. Consider those students who scored 90% or better in 4th grade. The streamlines and trajectory lines for those students are very similarly shaped for the class of 2011 and the class of 2012. This seems reasonable as the limited funds from ARI/AMI would not be likely spent on high-performing 5th graders and 8th graders. However, the *number* of students who populate the streamlines and trajectories is much higher for the class of 2012 than the class of 2011. Table 4.2 shows how the distribution of scores dramatically changes from the class of 2011 to the class of 2012. Among low-income students, the topmost trajectory (i.e. those students who scored between 90% and 100% in 4th grade) had a population increase of over 11,000 students for 11th grade, despite having almost identical scores on average to the previous cohort. This type of increase

Number of students in top trajectory	Better-Off		Low-Income	
	4th Grade	11th Grade	4th Grade	11th Grade
Class of 2011	35394	29760	22754	18708
Class of 2012	41892	35341	37045	30392
Number of students in bottom trajectory	Better-Off		Low-Income	
	4th Grade	11th Grade	4th Grade	11th Grade
Class of 2011	882	587	8112	4729
Class of 2012	493	263	6129	2015

Table 4.2: Comparison of trajectory populations between the classes of 2011 and 2012, as seen in Figure 4.14

cannot be attributed to population growth, which was approximately 2% per year for Texas and does not explain the depopulation of the under-performing trajectory lines [3].

4.3.3 Aftereffects

The longitudinal analysis of the TAKS test scores using our model highlighted a perturbation that turned out to the fingerprint of the ARI/AMI policy. This appears to have been a highly effective policy; students in the class of 2012 passed the high-stakes tests at a much greater rate than in previous cohorts, and the test scores increased at a statewide level that would have surpassed the gains predicted to be made by firing all the worst teachers in Texas and replacing them with superstar teachers. Moreover, the gains in test scores were not limited to just one year, as non-high-stakes testing years showed notable score increases as well. Finally, we found a large increase for the class of 2012 in the number of students that scored at the highest level in

any given year, with observed population increases for low-income students of over 60% more than in prior cohorts.

However, this policy was abruptly defunded for the 2009-10 school year (see Section 4.2). Why would the Texas Legislature shy away from a policy that had such great success in helping students to pass their high-stakes tests? It is true that Texas is funding a multitude of educational policies at any given time, and the budget must be stretched to accommodate all of them. It is possible that after a few years of SSI (although only six years of AMI), legislators wished to try something new to see if results might be improved further. On the other hand, reports on the effectiveness of the SSI program suggest that the success of the program was not fully understood.

In January 2009 an evaluation of the Student Success Initiative was presented to the 81st Texas Legislature, but it only discussed the educational impact of SSI on the 2006-07 school year [45]. At this point, the AMI component of ARI/AMI had only been in effect for four years. The report points out that for students that were identified as struggling in math at the beginning of the 2006-07 school year, 68% of those aided by ARI/AMI were assessed to be at grade level in math by the end of the year.[4] However, the report also says that the success rate of ARI/AMI students has remained relatively constant over time, at a rate of approximately two out of three students being able to

[4]Approximately 80% of those who were struggling in reading and/or math received the benefit of ARI/AMI funding; the rest used services not affiliated with ARI/AMI ($\sim 15\%$) or left the district ($\sim 5\%$).

raise their abilities up to grade level. Furthermore, the percentage of students who are identified as struggling at the *beginning* of each school year remained relatively unchanged, approximately 29% for reading and 20%-25% for math [45]. These percentages are determined by averaging over all grades K-7 for the 2006-07 school year; a breakdown of each grade level is listed in Table 4.3.[5] The process used to determine which students are considered struggling at the beginning of a school year is not transparent; the report defines "students who are struggling" for the purposes of ARI/AMI funding to be those who fail the first administration of the TAKS test, but also allows for struggling students to be identified through "diagnostic assessment tools" or classroom performance. Regardless, the fact that the percentage of struggling students was not decreasing at the beginning of each year suggests that score gains achieved in one year do not carry over to the next year, although the report itself notes that "this analysis was not longitudinal" [45] and "[i]t may be that the percentage of students identified as struggling remains constant as new students are identified as struggling" [45].

The legislators seemed to be looking for a continuous decrease in the percentage of struggling students. It was hoped that over time the number of students struggling in reading and math would be eventually reduced to zero, or close to it. However, our data shows that the conclusions of the SSI evaluation were accurate but not comprehensive. Score gains do persist from

[5]Note also in Table 4.3 that pre-ARI/AMI grades are not included, making it impossible to compare the percentage of struggling students from before ARI/AMI to afterwards.

Grade	K	1	2	3	4	5	6	7	K-7
2003-04	15%	17%	19%	26%	21%				20%
2004-05	17%	15%	18%	28%	27%	28%			22%
2005-06	15%	18%	21%	31%	30%	30%	24%		24%
2006-07	17%	20%	24%	33%	30%	30%	24%	25%	25%

Table 4.3: Reproduction of the table "Percentage of Students Identified as Struggling in Math, 2003-04 to 2006-07 School Years" [45]

one year to the next as shown in Figure 4.14, even in non-high-stakes years. The ARI/AMI program caused a net decrease in the number of struggling students which stayed nearly constant over time. A new steady state had been created that was an improvement over the last one, but it was only a steady state. Our methodology would have shown this. In the future, the techniques developed in this thesis may help to determine whether a policy like ARI/AMI is successful or not.

4.3.4 Longitudinal Limitations

The recommendations of the biennial evaluation included using longitudinal data to "...determine whether the accelerated instruction provided with ARI/AMI funds is sufficient to support students who are struggling in reading or mathematics, not only within the boundaries of one academic year, but over time as they progress through the education system" [45]. Why was this analysis not undertaken for the biennial report? One possible reason is that the data either had not been collected or had not been formatted in such a way that would allow for longitudinal analysis. As mentioned in Section 3.2,

the data files are separated according to their test administration, and much work was put into figuring out how to resolve problems relating to retests and duplicate test rows. It is unclear if the biennial report even had access to the full TAKS results; all the graphs and tables in the report cite their information from the 2006-07 eGrants Database Consolidated Reading Initiative Report and/or the 2005-06 ARI/AMI Final Evaluation Report, both from the Texas Education Agency.

Another possible explanation behind the lack of longitudinal analysis is that there simply was not enough data to draw concrete conclusions. Only four years of data on AMI had been collected for the report, and while this might be enough to give some qualitative estimate of the progress of students, it is not necessarily comprehensive. Similar to the contrast between streamline plots and trajectory plots, time is the ultimate cost of extracting enough data to conduct a full longitudinal analysis. Unfortunately, governments and legislatures do not usually have the benefit of waiting an indefinite period of time for all the data to be collected; often decisions must be made quickly using only the available data. The question of how much data is ultimately sufficient is one that will be addressed further in Section 5.2.1.

Chapter 5

Extensions

5.1 Utilizing Additional Data Sets

5.1.1 FERPA-Protected Data

Other states have collected vast quantities of data in relation to their own standardized tests. Notably, Washington and Michigan are two examples of states that have also developed educational research centers to allow access to FERPA-protected data [36]. Since we do not rely on scaled or vertical score scales, our model should be able to analyze test scores in other states as well. The challenges that face the implementation of our model are twofold. First, getting access to the data may involve contending with extensive bureaucratic hurdles. In Texas, access to FERPA-sensitive data requires research proposals that have been approved by the Joint Advisory Board (JAB) of the Texas Education Research Center [21]. These proposals are generally written in such a way to be very restrictive on what data may be accessed and on who may access it (hence, access to TAKS data does not confer access to STAAR data). However, the official website of the Texas Higher Education Coordinating Board mentions no meetings of the JAB since June 2009 [17]; our access

to the TAKS data was renewed in December 2011 and December 2013, but no online record of these meetings exists.[1] It is unclear how long a research group may have to wait in order to submit or renew a research proposal. Second, if we are able to clear these hurdles and gain access to the data, we will still need to learn a new data format (presuming that the new state's data will have a wholly different format than Texas's), understand all its idiosyncrasies, and possibly scrub and reformat it to produce a longitudinal data set. This is not necessarily a difficult process for someone who understands the code that was used to scrub the TAKS data, but it will take time. If the new state has similar security protocols as Texas (e.g. no remote access to the data files), this hypothetical situation would quickly become intractable. Obviously, the security of FERPA-protected data is a primary concern and careful steps should be taken to maintain it, but it seems that there should be some middle ground where potential research projects are not derailed because of distance.

There might be an alternative to the data that is being collected in service of No Child Left Behind. The University of Texas is one of the largest public school systems in the country, and has information about everything from students' grades to teacher evaluations. There is much potential in applying our model to this wealth of data, and concerns about security/distance could be largely alleviated by being affiliated with the university. This option is still being explored, but I feel it represents one of our best avenues of future

[1] An ERC advisory board meeting was held in January 2014, but there is no record of what was discussed or what proposals were reviewed [22].

research. Some questions we have considered include "Is a student more likely to stay in his or her major if the student has a great instructor for introductory courses?" and "How does the quality of lab courses affect grades in lectures, or vice versa?" The data set is not as large as that of the entire state of Texas, but there are more than enough students to ensure statistical rigor.

5.1.2 Course Instructor Survey Results

The University of Texas allows students to anonymously give feedback to their instructors on how they performed in each class. This is done through the use of Course Instructor Surveys (CIS) given at the end of the semester. These surveys consist of multiple-choice questions following the Likert-type scale, as well as a section for handwritten comments. Much of the information contained in these surveys is protected by FERPA, including the handwritten comments. However, the University of Texas has decided to allow the release of nine of the questionnaire items (see Table 5.1), provided there were enough students in the class to accommodate FERPA. These CIS results may be viewed through the University of Texas' secure web-portal, which restricts access to students, faculty members, or people who are otherwise officially associated with the University.

As a side project, I developed Python code that would automate the process of logging in to the secure web-portal, accessing the CIS results for each instructor, and collecting/processing the HTML code to create a database of the survey results. The data released through this method only went back

The course was well organized.
The instructor communicated information effectively.
The instructor showed interest in the progress of students.
The tests/assignments were usually graded and returned promptly.
The instructor made me feel free to ask questions, disagree, and express my ideas.
At this point in time, I feel that this course will be (or has already been) of value to me.
Overall, this instructor was:
Overall, this course was:
In my opinion, the workload in this class was:

Table 5.1: Nine CIS questions accessible through UT's web-portal

to the fall of 2005, but there were still around 10^5 web pages to collect. There were other gaps in the data from classes taught by teaching assistants (their CIS questionnaires do not include the main nine questions) or from classes where the instructor neglected to hand out surveys at all.

The original goal behind this side project was to look at how instructors were rated in the physics department, and possibly identify any areas that could be improved. The physics department generally has some of the lowest CIS scores at the University of Texas, and while there are a host of theories as to why that might be, I was only concerned about finding any trends that could suggest ways to improve our scores. Figure 5.1 compares the arrow plots of the CIS scores for the entire University of Texas and for just the physics department. While the CIS scores across the university are very uniform from semester to semester, the arrows in the physics plot experience large changes over time. No clear trend is evident in the physics plot.

However, I believe that the CIS data may be a candidate for our longitudinal model. Assuming that a professor's quality of teaching does not vary wildly over time, and that students have the ability to rate professors in a consistent manner, the CIS results could be considered semi-deterministic. We could generate flow plots and streamlines and look for persistent trends and perturbations in the system, hoping to identify things that work or areas that need improvement. We need to verify that the semi-deterministic hypothesis holds, and there are other considerations that go beyond our expertise (e.g. how often do students accurately rate their professors instead of just marking "Neutral" for every response?). Still, I believe this data set would be one way to extend our model.

5.2 Research Questions

5.2.1 Streamlines vs. Trajectories

During our development of streamline and trajectory plots, we were struck by their qualitative similarities. When they were applied to the TAKS data set and focused on the class of 2012, both types of plots showed peaks in 5th and 8th grade due to SSI, along with a large drop in 6th and 9th grade (though not so large a drop as to return the scores to the level of the class of 2011). However, the scores in 11th grade emphasized the difference between the two plotting styles. Figure 4.12 shows that all students are likely to score in a narrow band of 15-20 percentile points in 11th grade, no matter what score

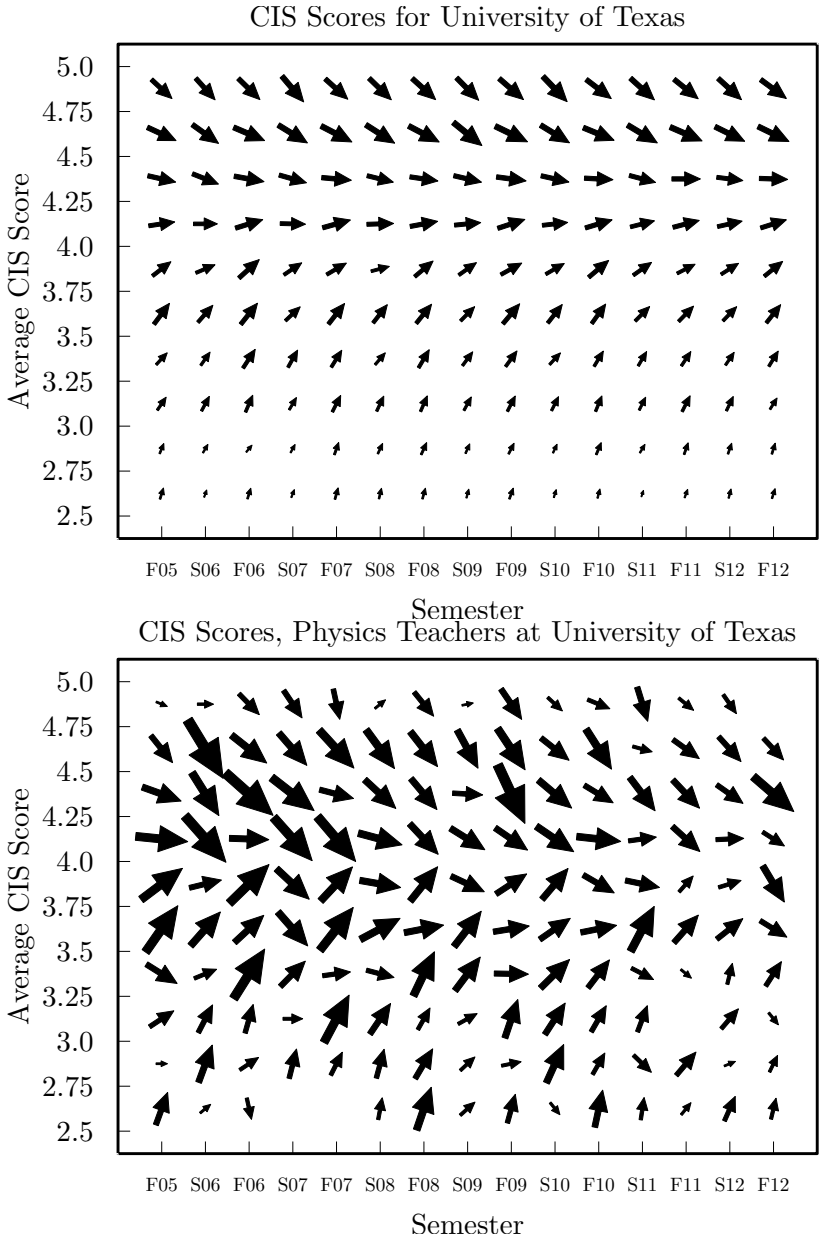

Figure 5.1: Arrow plots for the CIS scores of the University of Texas and of the physics department at UT. The sizes of the arrows have been enlarged in the physics plot to improve legibility. Each arrow shows on average how teachers' CIS scores changed from one semester to the next. Unlike the university-wide scores, the sizes and angles of the physics teachers' arrows are more erratically distributed.

they got in 3rd grade.[2] By comparison, if we actually follow the students from 4th to 11th grade in the trajectory plots of Figure 4.13, we see that students who score in the 90th percentile in 4th grade are likely to also score highly in 11th grade, or at least significantly higher than those students who started out in the 80th percentile. These observations are not unexpected; streamlines integrate over the flow pattern, and a casual look at Figures 4.4, 4.5, 4.6, and 4.7 suggests that there is a peak in the score distribution such that arrows above this peak will point downward, and arrows below it will point upward. In other words, we observe regression to the mean, and it is unsurprising that integration causes our streamlines to collapse.

Trajectories are more accurate than streamlines at showing how students' scores change on average over many years, as discussed in Section 2.4.4. Unfortunately, it took nine years to gather all the TAKS data associated with the class of 2012. Streamlines can show similar qualitative results, but they only require two years of data. Is there a way to get the best of both worlds, i.e. produce graphs that are reasonably accurate both quantitatively and qualitatively, but that do not take nine years to produce?

When constructing streamlines from a snapshot, we integrate over a flow pattern that was built by looking at two years of data, calculating the score change from one year to the next, and declaring that to be the average. In doing so, we throw away all prior information about those students. How-

[2]Separating the students by their economic status narrows the estimated range of 11th grade scores even further.

ever, it is possible that having an extra year of information could improve the predictive powers of our model. For example, we may say that students who score in the 80th percentile in 6th grade are most likely to score in the 70th percentile in 7th grade. But what is the most likely score for those students in 8th grade? What about those students that score in the 60th percentile in 6th grade and in the 70th percentile in 7th grade? Would we expect their scores to increase again, to regress, or to stay the same?

Preliminary research into this question suggests that it is more likely that students experience a negative score change after a positive one, or a positive score change after a negative one. It is less likely that students would experience increasing scores (or decreasing scores) in three consecutive years. Therefore, having flow plots where each arrow depends on multiple historical scores may result in patterns that more accurately reflect the true trajectories. An open research question is how to determine the number of years necessary in order for multiple-year streamline plots to be within a certain confidence interval of the trajectory plots, especially at later time steps (e.g. starting from 3rd graders in 2003, how many years of data do you need to effectively predict their scores in 11th grade?).

We have begun to establish a rigorous framework that would answer this question [15]. We adopt a Langevin equation framework [39] that takes the scores of each student to be a deterministic function of past scores plus a random component. This is similar to classical testing theory which submits that any student's test score is a combination of an underlying knowledge (or

"true knowledge") plus a random error term [23]. Consequently, a student's test score s_i may be represented by:

$$s_i = T_i + \xi_i$$

where T_i is the student's "true" score and ξ_i is the random error term.

Our Langevin framework does not presume that a student has an underlying knowledge state. Instead, we define $V(s_t, s_{t-1}; t)$ to be the deterministic function that predicts how a student's score will change based upon his or her two prior scores.[3] If we include a random variable ξ_t to represent noise in the system, then we have the following:

$$s_{t+1} = s_t + V(s_t, s_{t-1}; t) + \xi_t \tag{5.1}$$

We will assume ξ_t to be normally distributed such that its probability distribution is given by:

$$P(\xi_t) = \sqrt{\frac{1}{2\pi D}} e^{-\xi_t^2/2D} \tag{5.2}$$

where D is the variance of the distribution. In the event that the variance

[3]For now, we only examine the scenario where streamlines are determined using two years of history.

vanishes, $P(\xi_t) \to 0$ and Eq. 5.1 becomes completely deterministic. We will use the notation S_t to denote any scores that lie on these deterministic trajectories, so therefore $V(\mathsf{S}_t, \mathsf{S}_{t-1}; t)$ exactly represents the score change from one year to the next along a trajectory.

Therefore, to determine how quickly streamlines and trajectories diverge, we need to calculate the rate at which $v_{t,g,k}$ and $V(\mathsf{S}_t, \mathsf{S}_{t-1}; t)$ diverge. We will begin by adopting the convention of using σ_t (instead of s_t) to represent scores which will be integrated. This change of variable turns Eq. 5.1 into:

$$\sigma_{t+1} = \sigma_t + V(\sigma_t, \sigma_{t-1}) + \xi_t \tag{5.3}$$

We will suppress the t index in our function V from this point forward.

Next, we will calculate the probability of achieving a sequence of scores $\sigma_0, \sigma_1, \ldots, \sigma_T$ from the initial test to an arbitrary time T. This probability distribution will be essential for calculating expectation values. Begin by assuming that $V(\sigma_0, \sigma_{-1}) = V(\sigma_0)$; in other words, the change from σ_0 to σ_1 only depends on σ_0, but all other score changes will depend on two prior scores. Then:

$$
\begin{aligned}
\sigma_1 &= \sigma_0 + V(\sigma_0) + \xi_0 \\
\implies P(\sigma_0, \sigma_1) &= P(\sigma_0)P(\sigma_1|\sigma_0) \\
&= P(\sigma_0)P(\xi_0) \\
&= P(\sigma_0)\sqrt{\frac{1}{2\pi D}}e^{-\xi_0^2/2D}
\end{aligned}
$$

The assertion that $P(\sigma_1|\sigma_0) = P(\xi_0)$ is another way of saying that the probability of getting a score σ_1 given a prior score σ_0 is equal to the probability of getting the appropriate random term ξ_0 such that Eq. 5.3 is satisfied.

Also, we know that σ_0 and ξ_0 are completely independent of each other by definition. So we can draw a relation between between ξ_0 and σ_1:

$$
\begin{aligned}
P(\xi_0) &= \frac{P(\xi_0)P(\sigma_0)}{P(\sigma_0)} \\
&= P(\xi_0|\sigma_0) \\
\therefore P(\xi_0) &= P(\sigma_1) \qquad (5.4)
\end{aligned}
$$

Continuing the derivation:

$$\begin{aligned} \sigma_2 &= \sigma_1 + V(\sigma_1, \sigma_0) + \xi_1 \\ = P(\sigma_0, \sigma_1, \sigma_2) &= P(\sigma_0, \sigma_1) P(\sigma_2 | \sigma_0, \sigma_1) \\ &= P(\sigma_0, \sigma_1) P(\xi_1) \\ &= P(\sigma_0) \left(\sqrt{\frac{1}{2\pi D}} \right)^2 e^{-(\xi_0^2 + \xi_1^2)/2D} \end{aligned}$$

Using recursion, we have:

$$P(\sigma_0, \ldots, \sigma_T) = P(\sigma_0) \left(\sqrt{\frac{1}{2\pi D}} \right)^T \exp \left\{ -\sum_{i=0}^{T-1} \xi_i^2 / 2D \right\} \tag{5.5}$$

We now have a probability distribution for calculating the expectation values of score changes. Our velocity term $v_{t,g,k}$ can be written as the expectation value of $\sigma_{t+1} - \sigma_t$. More specifically, we have the expression:

$$v_{t,g,k} = \left\langle (\sigma_{t+1} - \sigma_t) \frac{\delta(\sigma_t - s_t)}{\langle \delta(\sigma_t - s_t) \rangle} \right\rangle \tag{5.6}$$

Let $P(s_t) = \langle \delta(\sigma_t - s_t) \rangle$ be the probability that the score at time t has value s_t; we include it in the denominator of Eq. 5.6 as a normalizing factor for the expectation value. Therefore, Eq. 5.6 can be expressed as the integral:

$$v_{t,g,k} = \int (\sigma_{t+1} - \sigma_t) \frac{\delta(\sigma_t - s_t)}{P(s_t)} P(\sigma_0, \ldots, \sigma_T)\, d\sigma_0 \cdots d\sigma_T \qquad (5.7)$$

We note that for all t' such that $t+1 < t' \leq T$, the only term that depends on $d\sigma_{t'}$ is $P(\sigma_0, \ldots, \sigma_{t'})$, and so we can calculate this integral (noting that Eq. 5.3 shows that $d\sigma_{t+1}/d\xi_t = 1$):

$$\begin{aligned}
\int P(\sigma_0, \ldots, \sigma_{t'})\, d\sigma_0 \cdots d\sigma_{t'} &= \int P(\sigma_0, \ldots, \sigma_{t'-1}) \sqrt{\frac{1}{2\pi D}} e^{-\xi_{t'-1}^2/2D} \\
&\quad \times d\sigma_0 \cdots d\sigma_{t'-1} d\sigma_{t'} \\
&= \int P(\sigma_0, \ldots, \sigma_{t'-1})\, d\sigma_0 \cdots d\sigma_{t'-1} \\
&\quad \times \int \sqrt{\frac{1}{2\pi D}} e^{-\xi_{t'-1}^2/2D} d\xi_{t'-1} \\
&= \int P(\sigma_0, \ldots, \sigma_{t'-1})\, d\sigma_0 \cdots d\sigma_{t'-1}
\end{aligned}$$

Therefore, integrating over all $d\sigma_{t+2} \cdots d\sigma_T$ reduces Eq. 5.7 to:

$$v_{t,g,k} = \int (\sigma_{t+1} - \sigma_t) \frac{\delta(\sigma_t - s_t)}{P(s_t)} P(\sigma_0, \ldots, \sigma_{t+1})\, d\sigma_0 \cdots d\sigma_{t+1}$$

For all t' such that $0 \leq t' < t-1$, we once again only have to integrate

$P(\sigma_0, \ldots, \sigma_{t'})$. From Eq. 5.5 and using the identity of Eq. 5.4, we have:

$$\begin{aligned}
\int P(\sigma_0, \ldots, \sigma_{t'}) d\sigma_0 &= \int P(\sigma_0) \left(\sqrt{\frac{1}{2\pi D}}\right)^{t'} \exp\left\{-\sum_{i=0}^{t'-1} \xi_i^2/2D\right\} d\sigma_0 \\
&= \left(\sqrt{\frac{1}{2\pi D}}\right)^{t'} \exp\left\{-\sum_{i=0}^{t'-1} \xi_i^2/2D\right\} \int P(\sigma_0) d\sigma_0 \\
&= \left(\sqrt{\frac{1}{2\pi D}}\right)^{t'} \exp\left\{-\sum_{i=0}^{t'-1} \xi_i^2/2D\right\} \\
&= \left[\sqrt{\frac{1}{2\pi D}} e^{-\xi_0^2/2D}\right] \left(\sqrt{\frac{1}{2\pi D}}\right)^{t'-1} \exp\left\{-\sum_{i=1}^{t'-1} \xi_i^2/2D\right\} \\
&= P(\xi_0) \left(\sqrt{\frac{1}{2\pi D}}\right)^{t'-1} \exp\left\{-\sum_{i=1}^{t'-1} \xi_i^2/2D\right\} \\
&= P(\sigma_1) \left(\sqrt{\frac{1}{2\pi D}}\right)^{t'-1} \exp\left\{-\sum_{i=1}^{t'-1} \xi_i^2/2D\right\} \\
&= P(\sigma_1, \ldots, \sigma_{t'})
\end{aligned}$$

Now we can integrate over all $d\sigma_0 \cdots d\sigma_{t-2}$ to reduce Eq. 5.7 to:

$$v_{t,g,k} = \int (\sigma_{t+1} - \sigma_t) \frac{\delta(\sigma_t - s_t)}{P(s_t)} P(\sigma_{t-1}, \sigma_t, \sigma_{t+1}) \, d\sigma_{t-1} d\sigma_t d\sigma_{t+1} \quad (5.8)$$

From Eqs. 5.3, 5.4, and 5.5, we can make a substitution to integrate

over $d\sigma_{t+1}$:

$$\begin{aligned}\int (\sigma_{t+1} - \sigma_t)\, P\,(\sigma_{t-1}, \sigma_t, \sigma_{t+1})\, d\sigma_{t+1} &= \int [V(\sigma_t, \sigma_{t-1}) + \xi_t] \\ &\quad \times [P\,(\sigma_{t-1}, \sigma_t)\, P\,(\xi_t)]\, d\xi_t \\ &= V(\sigma_t, \sigma_{t-1}) P\,(\sigma_{t-1}, \sigma_t) \int P\,(\xi_t)\, d\xi_t \\ &\quad + P\,(\sigma_{t-1}, \sigma_t) \int \xi_t P\,(\xi_t)\, d\xi_t \\ &= V(\sigma_t, \sigma_{t-1}) P\,(\sigma_{t-1}, \sigma_t)\end{aligned}$$

where the rightmost term disappears because $\langle \xi_t \rangle = 0$. Now we have:

$$v_{t,g,k} = \int V(\sigma_t, \sigma_{t-1}) \frac{\delta\,(\sigma_t - s_t)}{P\,(s_t)} P\,(\sigma_{t-1}, \sigma_t)\, d\sigma_{t-1} d\sigma_t$$

Integrating over $d\sigma_t$ to eliminate the delta function, we have:

$$v_{t,g,k} = \int V(s_t, \sigma_{t-1}) \frac{P\,(\sigma_{t-1}, s_t)}{P\,(s_t)} d\sigma_{t-1}$$

We are finally ready to calculate the rate at which $v_{t,g,k}$ and $V(\mathsf{S}_t, \mathsf{S}_{t-1})$

diverge. First, we make note of the following identity:

$$\begin{aligned} \int P\left(\sigma_{t-1}, s_t\right) d\sigma_{t-1} &= P\left(s_t\right) \\ \therefore \int \frac{P\left(\sigma_{t-1}, s_t\right)}{P\left(s_t\right)} d\sigma_{t-1} &= 1 \end{aligned}$$

Then we have the following:

$$\begin{aligned} v_{t,g,k} - V(s_t, \mathsf{S}_{t-1}) &= \int \frac{P\left(\sigma_{t-1}, s_t\right)}{P\left(s_t\right)} \\ &\quad \times \left[V(s_t, \sigma_{t-1}) - V(s_t, \mathsf{S}_{t-1})\right] d\sigma_{t-1} \end{aligned} \tag{5.9}$$

We now perform the following expansion of V to first order around S_{t-1}:

$$\begin{aligned} V\left(s_t, s'\right) &= V\left(s_t, \mathsf{S}_{t-1}\right) + \left(s' - \mathsf{S}_{t-1}\right) \left.\frac{\partial V}{\partial s'}\right|_{\mathsf{S}_{t-1}} \\ \therefore V(s_t, \sigma_{t-1}) - V(s_t, \mathsf{S}_{t-1}) &= \left(\sigma_{t-1} - \mathsf{S}_{t-1}\right) \left.\frac{\partial V}{\partial s'}\right|_{\mathsf{S}_{t-1}} \end{aligned}$$

Substituting into Eq. 5.9, we have:

$$\begin{aligned} v_{t,g,k} - V(s_t, \mathsf{S}_{t-1}) &= \int \frac{P\left(\sigma_{t-1}, s_t\right)}{P\left(s_t\right)} \left(\sigma_{t-1} - \mathsf{S}_{t-1}\right) \left.\frac{\partial V}{\partial s'}\right|_{\mathsf{S}_{t-1}} d\sigma_{t-1} \\ &= \left.\frac{\partial V}{\partial s'}\right|_{\mathsf{S}_{t-1}} \left[\overline{s}_{t-1}\left(s_t\right) - \mathsf{S}_{t-1}\right] \end{aligned}$$

where we have defined $\overline{s}_{t-1}(s_t)$ to be:

$$\overline{s}_{t-1}(s_t) = \int \sigma_{t-1} \frac{P(\sigma_{t-1}, s_t)}{P(s_t)} d\sigma_{t-1}$$

We interpret $\overline{s}_{t-1}(s_t)$ as follows: for those students who received a score s in year t, find their mean score from the year before. If $s_t > \overline{s}_{t-1}$, then the students must be some combination of those students whose true knowledge T is nearly equal to $\overline{s}_{t-1}$, and those students who have a lower true knowledge T but benefited from a positive random fluctuation E. By the logic of regression to the mean, this group has a lower mean score than what would have been calculated from pure deterministic reasoning; that is, $\overline{s}_{t-1}(s_t) < \mathsf{S}_{t-1}$. Similarly, if $s_t < \overline{s}_{t-1}$, then $\overline{s}_{t-1}(s_t) > \mathsf{S}_{t-1}$.

The term $\frac{\partial}{\partial s'} V(s_t, s')\big|_{\mathsf{S_{t-1}}}$ fixes the score s in year t and asks how the score in year $t+1$ will change as a function of $\mathsf{S_{t-1}}$. We make the assumption here that students who do better in year $t-1$ will perform better on average in year $t+1$, assuming identical scores in year t.[4] Therefore, $\frac{\partial}{\partial s'} V(s_t, s')\big|_{\mathsf{S_{t-1}}}$ should be positive, and the sign of $v_{t,g,k} - V(s_t, \mathsf{S}_{t-1})$ is determined solely by $\overline{s}_{t-1}(s_t) - \mathsf{S}_{t-1}$. If the deterministic $\mathsf{S_t}$ is greater than the recorded $\overline{s}_{t-1}$, then $v_{t,g,k} - V(s_t, \mathsf{S}_{t-1})$ is negative, and the trajectory will show higher scores in year $t+1$ than streamlines will. The reverse is true as well; if $\mathsf{S}_t > \overline{s}_{t-1}$, then the trajectories will be lower than the streamlines.

[4] Initial research has confirmed that this assumption is accurate on large scales, but more analysis is required.

In other words, the streamlines will experience regression to the mean far more quickly than trajectories will. This explains the figures in Section 4.3.2; the streamline plots fall into a narrow band of scores by 11th grade, but the trajectory plots stay spread out over a wide range of scores. Note that $V(s_t, s_{t-1})$ is just a deterministic function that relies on two years of prior scores. We have sketched out a formalism that shows why trajectories and streamlines diverge, but we could extend the formalism further and allow V to depend on three years of prior scores. Quantifying the differences between streamlines and trajectories will be a future topic of research in this field.

5.2.2 Transitioning Between Schools

As mentioned in Section 2.4.1, snapshot and cohort flow plots exhibit unusual behavior as students transition between 5th and 6th grade and between 8th and 9th grade. The arrows for these transitions tend to point sharply downward, and this effect is exacerbated for low-income students. Even if the better-off and low-income students have similar score distributions between 3rd and 5th grade, low-income students do much worse by comparison once they transition to 6th grade.

In Section 2.4.1, we suggested that this effect may be caused by the transition from elementary school to middle school, or from middle school to high school. Unfortunately, there is a problem with this idea; not all elementary schools are K-5 schools. In Texas we find K-6 schools, K-8 schools, and even schools that span between 1st and 11th grade [10]. There are approxi-

mately 30% as many K-6 schools as K-5 schools (see Section 2.4.1), and that does not begin to account for the other grade combinations. Is it possible to isolate the effect of transitioning from elementary to middle school?

We have some preliminary research that may be able to answer this question. We have created data files that disaggregate students based on what school they attended in a given grade. However, once we have disaggregated them in this way, we then create velocity grids for those students that span their entire educational career. For example, we might split students up depending on where they went to high school in 9th grade, but we still track them from 3rd to 11th grade (if the data is available).

Figure 5.2 shows an example of how we might use those data sets. Each school self-reports their grade levels to the TEA; we use the TEA's database from the 2010-2011 school year to figure out which campuses serve both 5th graders and 6th graders. By separating the students based on where they went to school in 6th grade, we can make cohort flow plots and compare the subsets directly. As Figure 5.2 shows, students who switch schools between 5th and 6th grade suffer large negative score changes on average when compared to students who stay in the same school for those years. This supports the suggestion that the transition between elementary and middle school is the reason for the sharp decline in our flow plots.

This idea could be explored further. The plots in Figure 5.2 are not disaggregated by free/reduced-lunch status, nor do they take into account whether the student populations differ at K-5 schools when compared to K-6

schools. Another issue is that of retention or dropouts; these graphs assume that if you went to a K-6 school in 6th grade, you were in the same school in 5th grade. That may be a reasonable assumption, but it might be less valid when considering the abnormally high retention/dropout rates in 9th grade. Much more data analysis is needed, but these preliminary results are promising.

5.2.3 Phase Transitions

In 2012, the first STAAR assessments were used for high-school students. The STAAR tests are fundamentally different from the TAKS tests in many respects, and the overall testing format was changed to favor subject tests over grade-specific tests. In May 2013, the Texas Legislature passed HB 5 that revises STAAR to only require five subject tests for high-school graduation (algebra I, biology, US history, English I, English II) instead of fifteen tests [4]. Such drastic changes in the STAAR format after only a couple years of implementation suggests that the entire system is unstable at the moment, and could change several more times as other legislation is introduced.

While the system is chaotic compared to the relatively constant testing format of TAKS, it does provide new avenues of research. Our model can show how students are adapting/coping with the change from TAKS to STAAR, and whether or not these changes are an overall positive or negative. By linking graduation data with other national tests like NAEP, we can judge if Texas students are more prepared to be college-ready. Since our model does not rely on scaled scores, we could make the assumption that raw test scores are

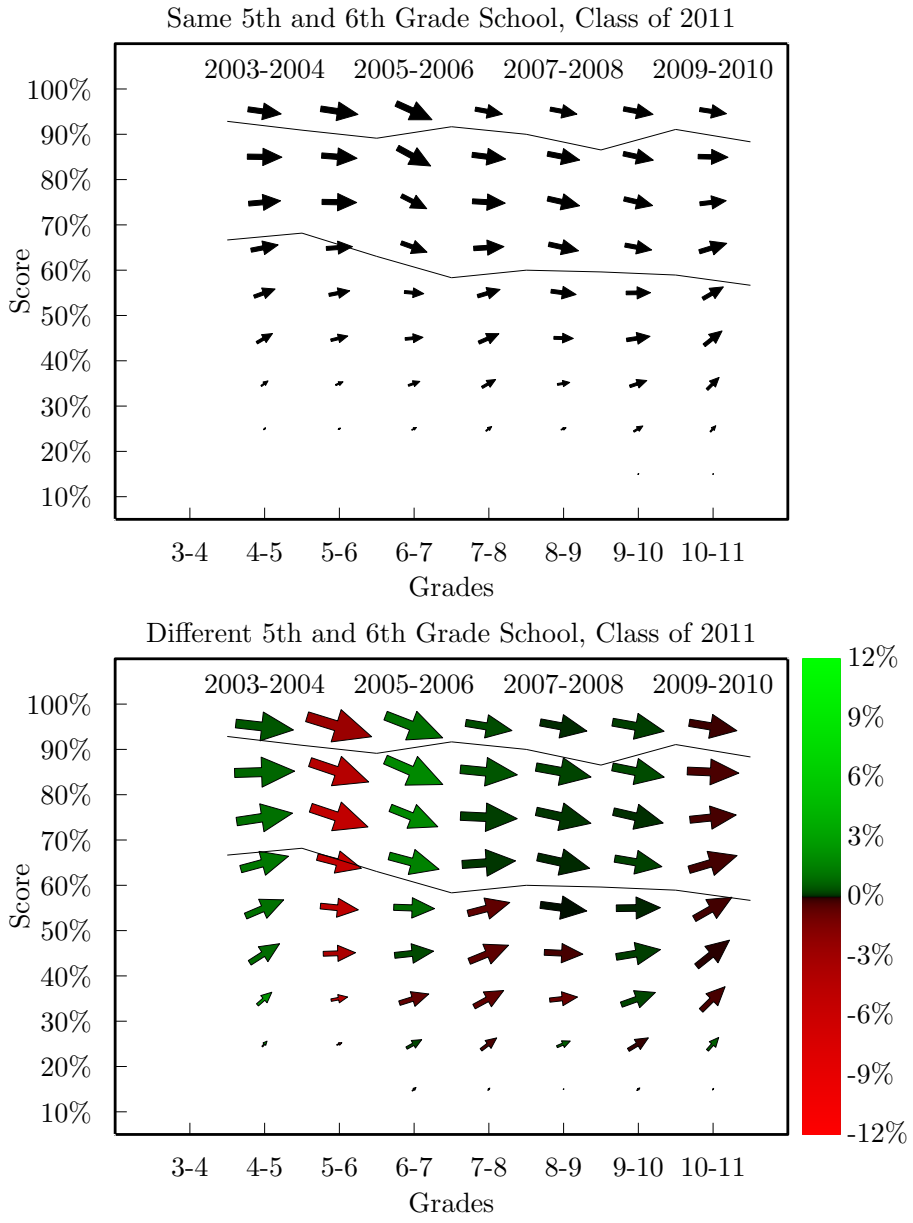

Figure 5.2: Comparison of students who attend different schools in 5th and 6th grade, and those that do not. Students who switch schools between 5th and 6th grade experience larger negative score gains than those who attend the same school in both grades.

equivalent between TAKS and STAAR, and see if something akin to a phase transition has occurred by looking for significant shifts in the flow patterns. There are many different topics of this nature that could be examined if we are allowed access to the data.

5.2.4 Characteristic Eigenvectors

The original Markov files described in Section 3.3 were tensors that answer the question, "For any two scores s and s', how many students scored s in year t and s' in year $t+1$?" For a fixed year and grade, the Markov files return a score-score matrix that looks like a single step in a Markov process. We have higher-order tensors, but the first-order Markov files eschew any student history beyond the two years under examination. We showed in Section 2.3 that if r_{ij} represents the probability that a student with score S_i in year A has a score S_j in year B, then $\hat{r} = \sum_i \sum_j r_{ij} \left|S_i\right\rangle \left\langle S_j\right|$ is a right stochastic matrix that describes the transition from A to B entirely. Since all the r_{ij} are probabilities, $\hat{r}$ is also a non-negative matrix. If $\hat{r}$ is representative of a large population, the Markov process that it describes is often irreducible (i.e. you can reach any score s' from any score s given enough steps) and aperiodic. The Perron-Frobenius theorem states that irreducible aperiodic non-negative matrices must have a stationary state eigenvector such that $\langle N|\,\hat{r} = \langle N|$. This leads to some interesting thought experiments; $\hat{r}$ is a representation of how a population of students performed over a two-year period, and $\langle N|$ is what that population would look like if they continued to perform in the same way

forever. Is it possible that the stationary states are in some way representative of a given population? If so, can it be applied at the campus or district level for comparative purposes?

There are some problems with trying to apply these stationary states to smaller populations. If we try to look at individual campuses, there may be only a few hundred students to populate the entire $\hat{r}$ matrix. In addition to the normal problems caused by FERPA masking, this also increases the chances that $\hat{r}$ has an absorbing state, or is otherwise reducible. This would mean that the Perron-Frobenius theorem no longer applies, or that the stationary state is nonsensical (i.e. a delta function). One solution might be to further coarse-grain the system to use fewer than 10 score bins, but that still may not be enough for some of the smallest high schools in Texas. For now, this topic is one of potential interest, but much more work is necessary to see if it is feasible.

Chapter 6

Conclusions

We have taken the model initially proposed by Marder and Bansal and extended it to be a much more powerful tool for visualizing longitudinal data. By returning to the fluid mechanical inspirations from which the model was derived, streamline and trajectory plots naturally emerged and showed different ways in which data might be presented to a large audience. Our sample system, the TAKS standardized test scores, ended up having a complex structure and an unexpected perturbation. The fact that the existence of the ARI/AMI program was discovered from our visualizations without any *a priori* knowledge or assumptions showcases one of the underlying strengths of this model; namely, that large-scale perturbations are easily detected and their effect on the population can be quickly ascertained.

The code that powers the data collection and data analysis has been rewritten several times and is now flexible enough to handle data from many different sources. It is designed to work independently of operating system and rely solely on the open source software Python instead of any proprietary software. As long as the data is semi-deterministic, longitudinal, and can be

initially displayed in a simple comma-separated values format, our methodology should be applicable.

This last part is important not only for general extensibility of our model, but also because the face of education is changing constantly. The STAAR standardized test was introduced in Texas for the 2011-12 school year, and it contains many revisions and changes from the way TAKS was administered. For example, STAAR is designed to include questions of increased rigorousness and difficulty, its science and math tests will include open-ended griddable items instead of being strictly multiple choice, and the tests in high school focus on specific subjects (e.g. geometry, U.S. history, chemistry) instead of generically discussing social studies or reading [7]. These changes may simply introduce perturbations into the score-grade continuum formed by the TAKS test, or they may represent an entire phase transition such that the two are not comparable. Without access to the STAAR data at the time of this dissertation, those questions are not answerable. However, by choosing to use raw score percentages instead of scaled scores, we have at least entertained the possibility that the gap between the two tests can be breached. Should STAAR continue to be the test of choice for Texas school systems, a new series of flow plots may emerge containing new insights. As long as data is being collected and being made available for this type of analysis, there will be more work to be done and more questions to be answered.

Appendix

The code for this project was written entirely in Python. Certain modules were used extensively in my research, most notably among them Numpy and Scipy. These modules are open source and contain large libraries of mathematical methods. Instead of having to build my own functions to interpolate or perform matrix operations, Numpy and Scipy provided already-optimized code to accomplish many tasks.

Other notable modules used in Python:

- *pickle*: Saves complex data structures in a binary format
- *csv*: Efficiently performs read/write operations on CSV files
- *lxml*: Creates and reads XML and HTML file structures
- *re*: Performs regular expression searches, substitutions, and text-related tasks
- *matplotlib*: Plotting software with a MATLAB-like interface

The graphs in this thesis were produced using the Ipe drawing editor. Ipe is free software that renders XML data structures in a Postscript or PDF format. It also accepts LaTeX code directly, for easy display of LaTeX mathematical expressions. Ipe was developed by Otfried Cheong, and his main page may be found at http://ipe7.sourceforge.net/.

The conclusions of this research do not necessarily reflect the opinions or official position of the Texas Education Agency, the Texas Higher Education Coordinating Board, Texas Workforce Commission, or the State of Texas.

Bibliography

[1] Index page for the ESEA flexibility page. US Department of Education, http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html. Retrieved March 2014.

[2] No Child Left Behind Act of 2001. Pub. L. no. 107-110, 115 Stat 1425, 2002.

[3] Table 1. Intercensal Estimates of the Resident Population for the United States, Regions, States, and Puerto Rico: April 1, 2000 to July 1, 2010 (ST-EST00INT-01), U.S. Census Bureau, Population Division, September 2011.

[4] 83d Leg., R.S. Tex. H.B. 5, 2013.

[5] Texas Education Agency. Texas Assessment of Academic Skills and college entrance examination performance trends in Texas. Technical report, Austin, TX, 2003. Policy Research Report No. 16 (Document No. GE04 601 01).

[6] Texas Education Agency. Texas Education Agency Growth Model Pilot Application for Adequate Yearly Progress determinations under the No Child Left Behind Act. http://goo.gl/XwgZoT, 2009.

[7] Texas Education Agency. A Comparison of Assessment Attributes TAKS to STAAR. http://goo.gl/OnbMw3, October 2010.

[8] Texas Education Agency. 2012 Expanded State Summary Tables - Performance. http://ritter.tea.state.tx.us/ayp/2012/summaries12_exp.pdf, 2012.

[9] Texas Education Agency. Final List of Campuses in School Improvement. http://goo.gl/R7oiy3, 2012.

[10] Texas Education Agency. 2010-11 Download of AEIS Data. http://goo.gl/pfF0DG, September 2013.

[11] Texas Education Agency. STAAR Resources. http://www.tea.state.tx.us/student.assessment/staar/, 2013.

[12] Texas Education Agency. TAKS Revised Information Booklets. http://www.tea.state.tx.us/student.assessment/taks/infobooks/, 2013.

[13] Texas Education Agency. Texas Essential Knowledge and Skills. http://www.tea.state.tx.us/index2.aspx?id=6148, 2013.

[14] Texas Education Agency. 2013-2014 Student Success Initiative Manual. http://goo.gl/g2v6zo, 2014.

[15] Anthony J. Bendinelli and M. Marder. Visualization of longitudinal student data. *Physical Review Special Topics - Physics Education Research*, 8:020119, 2012.

[16] Richard D. Bingham, John S. Heywood, and Sammis B. White. Evaluating schools and teachers based on student performance: Testing and alternative methodology. *Evaluation Review*, 15(2), 1991.

[17] Texas Higher Education Coordinating Board. All Meetings - Texas Higher Education Coordinating Board. `http://www.thecb.state.tx.us/apps/Events/AllMeetings.cfm`. Retrieved September 2013.

[18] Anthony S. Bryk and Stephen W. Raudenbush. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Sage Publications, 1992.

[19] Richard Buddin. Measuring teacher and school effectiveness at improving student achievement in Los Angeles elementary schools. Los Angeles Times, `http://goo.gl/pMdstv`, May 2011.

[20] Matthew Di Carlo. A 'Summary Opinion' of the Hoxby NYC Charter School Study. Albert Shanker Institute, `http://shankerblog.org/?p=3083`, July 2011.

[21] Texas Educational Research Center. Data Security. `http://goo.gl/DdTG1H`. Retrieved September 2013.

[22] Texas Educational Research Center. The Texas ERC at The University of Texas at Austin. `http://www.utaustinerc.org/`. Retrieved March 2014.

[23] Linda Crocker and James Algina. *Introduction to Classical and Modern Test Theory*. Wadsworth Group/Thomson Learning, 1986.

[24] Beck Evaluation and Testing Associates, Inc. Setting Student Performance Standards for the Texas Assessment of Knowledge and Skills (TAKS). `http://goo.gl/yUIOTP`, November 2002.

[25] Robert Gordon, Thomas J. Kane, and Douglas O. Staiger. Identifying effective teachers using performance on the job. `www.brookings.edu/views/papers/200604hamilton_1.pdf`, 2006. Retrieved November 2010.

[26] Eric A. Hanushek and Steven G. Rivkin. Generalizations about using value-added measures of teacher quality. *American Economic Review*, 100:267–271, 2010.

[27] Caroline M. Hoxby, Jenny Lee Kang, and Sonali Murarka. Technical Report: How New York City Charter Schools Affect Achievement. National Bureau of Economic Research, September 2009.

[28] Caroline M. Hoxby, Sonali Murarka, and Jenny Kang. How New York City's Charter Schools Affect Achievement, August 2009 Report. Second report in series. Cambridge, MA: New York City Charter Schools Evaluation Project, September 2009.

[29] Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley Publishing Company, Reading, Mass., 1968.

[30] Jon Lorence and Anthony Gary Dworkin. Elementary grade retention in Texas and reading achievement among racial groups: 1994-2002. *Review of Policy Research*, 23(5):999–1033, 2006.

[31] M. Marder and D. Bansal. Flow and diffusion of high-stakes test scores. *PNAS*, 106:17267–17270, 2009.

[32] William M. Mason, George Y. Wong, and Barbara Entwisle. Contextual analysis through the multilevel linear model. *Sociological Methodology*, 14:72–103, 1983.

[33] Daniel F. McCaffrey, J. R. Lockwood, Daniel M. Koretz, and Laura S. Hamilton. *Evaluating Value-Added Models for Teacher Accountability.* RAND Corporation, Santa Monica, CA, 2004.

[34] Yasuo Miyazaki and Stephen W. Raudenbush. Test for linkage of multiple cohorts in an accelerated longitudinal design. *Psychological Methods*, 5(1):44–63, 2000.

[35] US Department of Education. Race to the Top Program Executive Summary. Washington DC, `http://www2.ed.gov/programs/racetothetop/executive-summary.pdf`, 2009.

[36] National Forum on Education Statistics. Forum Guide to Supporting Data Access for Researchers: A State Education Agency Perspective. National Center for Education Statistics, Washington DC, `nces.ed.gov/pubs2012/2012809.pdf`, 2012.

[37] Sarah E. Peterson, James S. DeGracie, and Carol R. Ayabe. A longitudinal study of the effects of retention/promotion on academic achievement. *American Educational Research Journal*, 24(1):107–118, 1987.

[38] Georg Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests.* University of Chicago Press, Chicago, 1980.

[39] Frederick Reif. *Fundamentals of Statistical and Thermal Physics.* McGraw Hill, 1965.

[40] Melissa Roderick and Jenny Nagaoka. Retention Under Chicago's High-Stakes Testing Program: Helpful, Harmful, or Harmless? *Educational Evaluation and Policy Analysis*, 27(4):309–340, 2005.

[41] W. L. Sanders and J. C. Rivers. Cumulative and residual effects of teachers on future student academic achievement: Research progress report. Technical report, Value-Added Research and Assessment Center, Knoxville, TN: University of Tennessee, 1996.

[42] William L. Sanders. Value-added assessment from student achievement data: Opportunities and hurdles. *Journal of Personnel Evaluation in Education*, 14:329–339, 2000.

[43] William L. Sanders and Sandra P. Horn. The Tennessee Value-Added Assessment System (TVAAS): Mixed-Model Methodology in Educational Assessment. *Journal of Personnel Evaluation in Education*, 8:299–311, 1994.

[44] Texas Education Agency. Texas Education Agency – Technical Digest 2007-2008. http://bit.ly/PUkI0e, 2008.

[45] Texas Education Agency. The Student Success Initiative: An Evaluation Report. Austin, TX, http://goo.gl/dYzC4P, 2009.

[46] Texas Education Agency. 2011 TAKS Data File Format. http://goo.gl/b8NnUo, 2011.

[47] Texas Education Agency. Texas Education Agency - 2012-2013 Schools in Need of Improvement - Campus Level. http://goo.gl/iG68GX, 2013.

[48] Texas Education Agency, Department of Assessment and Accountability, Division of Performance Reporting. 2012 Adequate Yearly Progress (AYP) Guide. http://goo.gl/KrH0YG, June 2012.

[49] Texas Education Agency, University of Texas at Dallas Education Research Center, Gibson Consulting, and Learning Points Associates an affiliate of American Institutes for Research. Student Success Initiative: Consolidated Report. Austin, TX, http://goo.gl/zQ0ym9, 2010.

[50] U.S. Department of Education. Mapping State Proficiency Standards Onto the NAEP Scales: Variation and Change in State Standards for Reading and Mathematics, 2005-2009, 2009. http://nces.ed.gov/nationsreportcard/pdf/studies/2011458.pdf, Retrieved July 2013.

[51] Benjamin D. Wright. Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems*, pages 85–101.

[52] S.P. Wright, S.P. Horn, and W.L. Sanders. Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, 11:57–67, 1997.

[53] Nailing Xia and Sheila Nataraj Kirby. *Retaining Students in Grade: A Literature Review of the Effects of Retention on Students' Academic and Nonacademic Outcomes.* RAND Corporation, Santa Monica, CA, 2009.

Vita

Anthony James Bendinelli attended St. Charles Preparatory School, Columbus, Ohio. In 2003 he attended the University of Notre Dame in South Bend, Indiana. During the 2005-2006 school year, he attended New College at Oxford University, UK. Upon graduating from Notre Dame in 2007 with a Bachelor of Science (Physics/Honors Mathematics), he immediately entered the Graduate School at the University of Texas at Austin.

Permanent address: 11411 Research Blvd. #123
Austin, TX 78759

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.